SEQUENTIAL INTERVAL ESTIMATION FOR BERNOULLI TRIALS

A Dissertation
Presented to
The Academic Faculty

By

Tony Yaacoub

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Industrial Engineering in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2018

SEQUENTIAL INTERVAL ESTIMATION FOR BERNOULLI TRIALS

Approved by:

Dr. David Goldsman, Advisor
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yajun Mei, Advisor
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Dr. George V. Moustakides
Department of Computer Science
Rutgers University

Dr. Branislav Vidakovic
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Christos Alexopoulos
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: July 20, 2018

To my parents, Hiame and Raefat.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincerest appreciation to my family for their unconditional love, sacrifice, and support. Without them, I would not even dream of being where I am today.

I would also like to express my deepest gratitude to my advisors, Dr. Dave Goldsman and Dr. Yajun Mei; and to my research mentor and collaborator, Dr. George Moustakides: I really appreciate your patience, guidance, support, and friendship. I certainly would not be here without you.

Special thanks to my committee members, Dr. Brani Vidakovic and Dr. Christos Alexopoulos: Thank you so much for graciously serving on my dissertation committee, and thank you for your helpful feedback and support. You are amazing!

To my friends in ISyE: Thank you for making these past four years so memorable and enjoyable. My warm thanks go to: German Schnaidt, Timur Tankayev, Mina Georgieva, Idil Arşik, Damian Reyes, Luke Marshall, Alfredo Torrico, Asteroide Santana, Denise Batista, Chih-Li Sung, Şeyma Güven-Koçak, Nayeon Kim, Chen Feng, Reem Khir, Daniela Hurtado, Yassin Ridouane, Şeyma Gürkan, Ian Herszterg, Catharina Hollauer, Marianna De Almeida Costa, Matias Siebert, Bernardita Rios, Adrian Rivera, Andew El Haber, Jeff Pavelka, Brian Kues, Daniel Zink, Mathias Klapp, Alvaro Lorca, Jikai Zou, Ben Johnson, Xiaolei Fang, Murat Yildirim, and so many others.

Thank you to some special people who have been family throughout these ten years while being away from home: Jeff and Tina Cooper; and Jerry and Lydia Murphy.

Finally, thank you to my collaborator, Dr. Robert Sargent; my undergraduate advisor, Dr. Elliot Krop; and my master's advisor, Dr. Charles Champ.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Interval estimation of a binomial proportion is one of the most-basic problems in statistics with many important real-world applications. Some classical applications include estimation of the prevalence of a rare disease and accuracy assessment in remote sensing. In these applications, the sample size is fixed beforehand, and a confidence interval for the proportion is obtained. However, in many modern applications, sampling is especially costly and time consuming, e.g., estimating the customer click-through probability in online marketing campaigns and estimating the probability that a stochastic system satisfies a specific property as in Statistical Model Checking. Because these applications tend to require extensive time and cost, it is advantageous to reduce the sample size while simultaneously assuring satisfactory quality (coverage) levels for the corresponding interval estimates. The sequential version of the interval estimation aims at the latter goal by allowing the sample size to be random and, in particular, formulating a stopping time controlled by the observations themselves. The literature focusing on the sequential setup of the problem is limited compared to its fixed sample-size counterpart, and sampling procedure optimality has not been established in the literature. The work in this thesis aims to extend the body of knowledge on the topic of sequential interval estimation for Bernoulli trials, addressing both the theoretical and practical concerns.

In the first part of this thesis, we propose an optimal sequential methodology for obtaining fixed-width confidence intervals for a binomial proportion when prior knowledge of the proportion is available. We assume that there exists a prior distribution for the binomial proportion, and our goal is to minimize the expected number of samples while guaranteeing that the coverage probability is at least a specified nominal coverage probability level. We demonstrate our stopping time is always bounded from above and below; we will need to first accumulate a sufficient amount of information before we start applying our stopping rule, and our stopping time will always terminate in finite time. We also compare our

method with the optimum fixed-sample-size procedure as well as with existing alternative sequential schemes.

In the second part of this thesis, we propose a two-stage sequential method for obtaining tandem-width confidence intervals for a binomial proportion when no prior knowledge of the proportion is available and when it is desired to have a computationally efficient method. By tandem-width, we mean that the half-width of the confidence interval of the proportion is not fixed beforehand; it is instead required to satisfy two different upper bounds depending on the underlying value of the binomial proportion. To tackle this problem, we propose a simple but useful sequential method for obtaining fixed-width confidence intervals for the binomial proportion based on the minimax estimator of the binomial proportion.

Finally, we extend the idea for Bernoulli distributions in the first part of this thesis to interval estimation for arbitrary distributions, with an alternative optimality formulation. Here, we propose a conditional cost alternative formulation to circumvent certain analytical/computational difficulties. Specifically, we assume that an independent and identically distributed process is observed sequentially with its common probability density function having a random parameter that must be estimated. We follow a semi-Bayesian approach where we assign cost to the pair (estimator, true parameter), and our goal is to minimize the average sample size guaranteeing at the same time an average cost below some prescribed level. For a variety of examples, we compare our method with the optimum fixed-sample-size and other existing sequential schemes.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Statistical Inference

Making inference about unknown population parameters was the prime motivation for the development of statistics as a science, and continues to be at the forefront of almost all statistical applications and research. In statistical theory, statistical inference has generally been divided in two main categories, namely, *estimation* and *testing*. In estimation, the objective is to estimate an unknown population parameter based on some observed data. In testing (or hypothesis testing), an attempt is made to validate or invalidate a hypothesis about a specific value of an unknown population parameter based on observed data. The area of estimation may lend itself to two subdivisions: *point estimation* and *interval estimation*. In point estimation, the objective is to estimate an unknown population parameter with a single value based on the observed data. The behavior of a predefined loss function can be often described by the precision and accuracy of the point estimate. For instance, under the square error loss function, this precision is measured by the mean squared error (MSE), where an estimator is better if it has a lower MSE. On the other hand, in interval estimation, the objective is to obtain a range of values within which, we believe, the true population parameter lies with high coverage probability. In interval estimation, the width of the interval and the confidence coefficient jointly provide a sense of precision and accuracy. We note that these seemingly different criteria may be unified under appropriately defined loss functions. This thesis is particularly concerned with developing novel methods for interval estimation of a population parameter, such as the binomial proportion, under a sequential setting. The next section formulates the interval estimation problem under the fixed-sample-size and sequential settings, highlighting the differences between the two approaches.

1.2 Interval Estimation

Assume that we observe a sequence of independent and identically distributed (i.i.d.) random variables $\{X_t\}$'s over times $t = 1, 2, \ldots, T$, with a common probability density function (p.d.f.) $\mathsf{f}(x|\theta)$ and parameter $\theta \in \Theta$ is considered random. For an estimator $\hat{\theta}_T$ of θ based on the accumulated observations $\{X_1, X_2, \ldots, X_T\}$, consider a $100(1-\alpha)\%$ confidence interval (CI) of width $2h$ such that $[\hat{\theta}_T - h, \hat{\theta}_T + h]$ has a confidence level or coverage probability (CP) $1-\alpha$. Equivalently,

$$\mathsf{P}_\theta(\theta \in [\hat{\theta}_T - h, \hat{\theta}_T + h]) = 1 - \alpha, \text{ for all } \theta \in \Theta, \tag{1.1}$$

where $\mathsf{P}_\theta(\cdot)$ denotes the probability *given* θ. Note that $\hat{\theta}_T$ may be a biased estimator of θ, but what matters is that

$$\mathsf{P}_\theta(|\hat{\theta}_T - \theta| \leq h) = 1 - \alpha.$$

In most applications, achieving a $1-\alpha$ CP throughout the range of θ is impossible. Thus, two approaches are used: (1) obtain the mean coverage probability (averaged over $\theta \in \Theta$) and set this mean CP to be approximately equal to the nominal coverage probability $1-\alpha$; or (2) guarantee that the coverage probability is at least the nominal $1-\alpha$ for every value of $\theta \in \Theta$. The latter is preferred in most applications, especially in clinical trials, as accuracy is very important to be guaranteed regardless of the true value of the unknown population parameter θ. Thus, in our work, we focus on

$$\mathsf{P}_\theta(|\hat{\theta}_T - \theta| \leq h) \geq 1 - \alpha, \text{ for all } \theta \in \Theta. \tag{1.2}$$

To distinguish between the fixed-sample-size and sequential methodologies, we refer back to (1.2). In the fixed-sample-size setting, T is fixed, and we are interested in calculating the half-width h of the CI such that we achieve at least the nominal $1 - \alpha$ CP. Notice that in this case, the sample size T is known to the experimenter. In the sequential setting, however, an experimenter gathers information regarding an unknown parameter by observing random samples in successive steps. One may take one observation at-a-time or a few at-a-time, but a common characteristic among such sampling designs is that the total number of observations collected at termination is a positive integer-valued random variable T, a stopping time. For the CI setting in (1.2), $h = h(\theta)$ is fixed beforehand and T is a random stopping time controlled by the observations themselves. It is important to note here that in the sequential setting, h is fixed in the sense that it is not necessarily constant, but the functional form as a function of θ is fixed. For instance, fixed-width CIs have a constant h, whereas relative-width CIs have a half-width $h = \eta\theta$, where η is a fixed constant between 0 and 1.

1.3 Why Sequential?

Before justifying the importance of the sequential setting in interval estimation, we first provide a brief history of sequential analysis as given in [1]. According to [1, pp. 1–2], the work of Mahalanobis in [2] on estimating the acreage of jute crop in the entire state of Bengal is considered by many, including Abraham Wald, as the forerunner of sequential analysis. Wald and his collaborators, then, systematically developed theory and methodology of sequential tests in the early 1940s to reduce the number of sampling inspections without compromising the reliability of the terminal decisions, leading to Wald's [3] classical book on sequential analysis and the famous Sequential Probability Ratio Test (SPRT).

The advantage of sequential analysis is that it can arrive at a decision much sooner and with substantially fewer observations than equally reliable test procedures based on a predetermined number of observations. The methodology's immense value was immediately recognized during that time, and its use was 'classified' and restricted to wartime research and procedures in response to demands for efficient testing of anti-aircraft gunnery and other weapons during World War II. In 1945, the work was released to the public and has since revolutionized many aspects of statistical practice.

Methodological researchers caught on and began applying sequential analysis to solve a wide range of practical problems from inventory, queuing, reliability, quality control, design of experiments, and multiple comparisons, to name a few. In the 1960s through the 1970s, researchers in clinical trials realized the relevance of emerging adaptive designs and optimal stopping rules. The area of clinical trials continues to be a success story for the sequential methodology.

In terms of interval estimation, sequential methods are generally more favorable than their fixed-sample-size alternative because of the ability to achieve the same accuracy while requiring fewer samples. Furthermore, there are many problems that cannot be solved by any fixed-sample-size methodology. For instance, there exists no fixed-sample-size methodology that can deliver a preassigned estimation error or half-width h such that (1.2) is satisfied, see, e.g., [1, pp. 16–17]. In the next section, we continue our discussion of the sequential interval estimation problem and its advantages over the fixed-sample-size counterpart by concentrating specifically on Bernoulli trials.

1.4 Bernoulli Trials

Interval estimation of a binomial proportion θ is one of the most-basic problems in statistics, with many important real-world applications. Some classical applications include interval estimation of the prevalence of a rare disease [4]; interval estimation of the overall response rate in clinical trials [5]; and accuracy assessment in remote sensing [6]. In these applications, the sample size is fixed in advance, and a confidence interval for θ is obtained. There is an extensive bibliography regarding derivations of confidence intervals for θ when the sample size is fixed. Perhaps the most-widely known in this category is Wald's interval, which takes the form

$$\hat{\theta}_T \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}_T(1-\hat{\theta}_T)}{T}},$$

where T is the fixed sample size, $1-\alpha$ expresses the desired coverage probability, $\hat{\theta}_T$ is the sample mean of θ and $z_{\alpha/2}$ satisfies $Q(z_{\alpha/2}) = \frac{\alpha}{2}$ with $Q(x)$ denoting the complementary cumulative density function (cdf) of a standard $N(0,1)$ Gaussian random variable. This confidence interval is derived based on the asymptotic normality of $\hat{\theta}_T$ and, therefore, exhibits poor behavior when $T\theta(1-\theta)$ is small [7–10]. Several efforts to improve Wald's classical method are reported in [7, 11–15]. There are also Bayesian-based techniques [8, 16, 17] while [7–10, 18] give interesting surveys that evaluate the relative performance of the above methods. Finally, we mention that [19] provides explicit formulas for the required sample size that can guarantee a prescribed coverage probability for the Clopper-Pearson [12] method.

In many modern applications, sampling observations is costly and time consuming. Therefore, there is a desire to limit the sampling size without compromising the quality of the interval estimate. For instance, in automatic fraud detection in finance, one needs to manually go through the "suspect" financial transactions that are automatically detected as fraudulent by some machine learning or other computer algorithm. Since the manual process is expensive in terms of labor and cost, it is desirable to quickly estimate, with high confidence, what percentage of the suspect transactions are truly fraudulent. A different motivating application is in Statistical Model Checking, where with an approximate verification method, one overcomes the state space explosion problem for probabilistic systems by the use of Monte Carlo simulations. Given an executable stochastic system, we seek to verify that the system satisfies a particular property via simulation; and we desire to estimate the probability θ by which the system actually satisfies the property in question. The goal is to estimate θ within acceptable margins of error and confidence (see [20] and references therein). Because Monte Carlo simulations very often tend to require extensive time and computing power, it is advantageous to reduce their number assuring, at the same time, satisfactory quality levels for the corresponding estimate.

The sequential version of the interval estimation aims exactly at reducing the sample size by selecting it to be random and, in particular, a stopping time controlled by the observations themselves. The literature focusing on the sequential setup of the problem is limited compared to its fixed sample-size counterpart (see [21–23]). However, none of these articles is able to claim optimality of their corresponding schemes in any sense.

1.5 Main Conrtibutions

In this thesis, we make three contributions to sequential CI research. First, as noted previously, none of the articles that focus on sequential fixed-width CI for a binomial proportion θ is able to claim optimality of their schemes. Thus, our first contribution is developing an optimal sequential interval method for θ, with the quality of the estimate expressed through the coverage probability. In this case, prior knowledge of θ is needed, as we need to predefine our fixed half-width h beforehand. This introduces another obstacle: if no prior knowledge of θ is available, a fixed h for all $\theta \in (0, 1)$ is problematic. For instance, if the true θ were actually close to 0 or 1, then a half-width $h = 0.1$, say, is meaningless. Thus, when we have no prior knowledge of θ and desire a computationally efficient method, our second contribution is developing a tandem-width CI that satisfies two different upper bounds, h_0 and h_1, depending on the true value of θ. The third contribution is presenting an alternative optimality formulation that circumvents the computational complexity of the optimal formulation for sequential fixed-width CIs that is known to be generally analytically intractable when prior knowledge of θ is available. This alternative optimality formulation is extended to sequential fixed-width interval estimation of a parameter of interest of any arbitrary distribution.

Below we present the details of our contributions covered in each chapter of this thesis, with motivations for each chapter along with a preview of the results found within.

- **Optimal Stopping for Interval Estimation in Bernoulli Trials.** In Chapter 2, we propose an optimal sequential methodology for obtaining confidence intervals for a binomial proportion θ. Assuming that an i.i.d. sequence of Bernoulli(θ) trials is observed sequentially, we are interested in designing (i) a stopping time T that will decide the best time to stop sampling the process, and (ii) an optimum estimator $\hat{\theta}_T$ that will provide the optimum center of the interval estimate of θ. We follow a semi-Bayesian approach, where we assume that there exists a prior distribution for θ, and our goal is to minimize the average number of samples while we guarantee a minimal specified coverage probability level. The solution is obtained by applying standard optimal stopping theory and computing the optimum pair $(T, \hat{\theta}_T)$ numerically. Regarding the optimum stopping time component T, we demonstrate that it enjoys certain very interesting characteristics not commonly encountered in solutions of other classical optimal stopping problems. In particular, we prove that, for a particular prior (Beta density), the optimum stopping time is always bounded from above and below; we need to first accumulate a sufficient amount of information before we start applying our stopping rule, and our stopping time will always terminate before some finite deterministic time. We also conjecture that these properties are present with any prior. Finally, we compare our method with the optimum fixed-sample-size procedure as well as with existing alternative sequential schemes.

- **Tandem-width Sequential Confidence Intervals for a Bernoulli Proportion.** In Chapter 3, we propose a two-stage sequential method for obtaining tandem-width confidence intervals for a Bernoulli proportion θ. By tandem-width, we mean that the half-width h of the $100(1-\alpha)\%$ CI is not fixed beforehand; it is instead required to satisfy two different upper bounds, h_0 and h_1, depending on the values of θ. To tackle this problem, we propose a simple but useful sequential method for obtaining fixed-width confidence intervals for θ based on the minimax estimator of θ. We observe independent and identically distributed Bernoulli(θ) trials sequentially, and for some fixed half-width $h = h_0$ or $h = h_1$, we develop a stopping time T such that the resulting confidence interval for θ, $[\hat{\theta}_T - h, \hat{\theta}_T + h]$, covers the parameter with confidence at least $100(1-\alpha)\%$, where $\hat{\theta}_T$ is the maximum likelihood estimate of θ at time T. Furthermore, we derive certain theoretical properties of our proposed method, compare the tandem-width procedure with the proposed fixed-width alternative, and compare the performance of our fixed-width procedure with existing alternative sequential schemes.

- **Sequential Parameter Interval Estimation based on Conditional Cost.** In Chapter 4, we extend the idea for Bernoulli distributions in Chapter 2 to interval estimation for arbitrary distributions, though we use an alternative optimality formulation. Note that the optimal formulation in Chapter 2 presents computational challenges and is generally analytically intractable. Here, we propose a conditional cost alternative formulation that sacrifices part of the optimal performance to circumvent the analytical/computational difficulties. Specifically, we assume that an i.i.d. process is observed sequentially with its probability density function having a random parameter that must be estimated. We are interested in designing a stopping time that will decide the best moment to stop sampling the process and an estimator that will use the acquired samples in order to provide the desired estimate. We follow a semi-Bayesian approach where we assign cost to the pair (estimator, true parameter) and our goal is to minimize the average sample size guaranteeing at the same time an average cost below some prescribed level. For our analysis we adopt a conditional average cost that leads to a considerable simplification in the sequential estimation problem, otherwise known to be analytically intractable. For a variety of examples, we compare our method with the optimum fixed-sample-size and other existing sequential schemes.

CHAPTER 2

OPTIMAL STOPPING FOR INTERVAL ESTIMATION IN BERNOULLI TRIALS

2.1 Introduction

As mentioned in Chapter 1, the literature focusing on the sequential setup of the problem is limited compared to its fixed sample-size counterpart (see [21–23]). Furthermore, none of these articles is able to claim optimality of their corresponding schemes in any sense. The objective of our current work in this chapter is to offer optimum *sequential* methods for interval estimation of θ, with the quality of the estimate expressed through the *coverage probability*. In addition to deriving the optimum scheme, we will also demonstrate some very uncommon but highly interesting properties of the optimum solution. These properties are not encountered in optimum sequential schemes derived for other well known sequential problems (i.e. sequential hypothesis testing). We must also add that our methodology exhibits similarities with the work developed in [24]. However, the focus in [24] is on the actual estimate of θ with the adopted criterion being a variation of the classical mean square error. In our work, as we pointed out, we focus on confidence intervals and coverage probabilities; and, as it turns out, this difference makes our derivations and proofs far more complicated, requiring original analytical methodology. This becomes particularly apparent when we attempt to establish the validity of the unique properties, mentioned before, that characterize our optimum solution.

The remainder of this chapter is organized as follows. In Section 2.2 we discuss our proposed framework for interval estimation for θ and propose a well-defined optimization problem and discuss its general solution. In Section 2.3 we focus on the computational aspects of the optimum scheme and the unique properties that they characterize it. In Section 2.4 we compare the proposed scheme against the fixed-sample-size and two existing sequential methods in the literature. Finally, Section 2.5 contains our conclusions, and Section 2.6 contains the proofs of the major lemmas and theorems in this chapter.

2.2 Proposed Framework

We observe sequentially an i.i.d. process $X_1, X_2, \ldots$ of Bernoulli random variables with $X_t \in \{0,1\}$ and $\mathsf{P}(X_t = 1) = \theta = 1 - \mathsf{P}(X_t = 0), \theta \in [0,1]$. The goal is to provide a confidence interval for θ. We are interested in confidence intervals of fixed width equal to $2h$ for some pre-specified $h \in (0, \frac{1}{2})$. We would also like our scheme to be able to guarantee a coverage probability equal to $1-\alpha$, where $\alpha \in (0,1)$ is given. Our scheme consists of a pair $(T, \hat{\theta}_T)$, that is, a stopping time T and a *mid-point estimator*[1] $\hat{\theta}_T$, where T is adapted to the observation history (filtration generated by the observations) and $\hat{\theta}_T$ is a function of the observations accumulated up to the time of stopping T. We would like to solve the following constrained optimization problem for the optimum pair

$$\inf_{T,\hat{\theta}_T} \mathsf{E}_\theta[T], \text{ subject to: } \mathsf{P}_\theta(|\hat{\theta}_T - \theta| > h) \leq \alpha, \tag{2.1}$$

where the desired interval estimate is $[\hat{\theta}_T - h, \hat{\theta}_T + h]$ (with the two ends cropped at 0 and 1, respectively, whenever they exceed the two limits) and where $\mathsf{P}_\theta(\cdot)$ and $\mathsf{E}_\theta[\cdot]$ denote probability and expectation for *given* θ.

Although (2.1) seems as the ideal formulation, it unfortunately targets an infeasible goal. We note that we are asking for the pair $(T, \hat{\theta}_T)$ to minimize the average number of samples *for every value* of the parameter θ. In other words, we want our scheme to enjoy a *uniform* optimality property over all θ, a requirement which is impossible to satisfy. In order to be able to find a solution that has a well-defined form of optimality, we adopt a semi-Bayesian approach[2] and assume that a prior $\pi(\theta)$ for θ is available. This allows for

[1]The estimate $\hat{\theta}_T$ does not have the meaning of a classical parameter estimator. It is the mid-point of the confidence interval $[\hat{\theta}_T - h, \hat{\theta}_T + h]$ and does not necessarily constitute an efficient estimate of θ.

[2]The term "semi-Bayesian" is used because our setup involves two different components where one is optimized while the other is constrained, unlike full-Bayesian approaches that combine all terms into a single

the following modification of the previous constrained optimization

$$\inf_{T,\hat{\theta}_T} \mathsf{E}[T], \text{ subject to: } \mathsf{P}(|\hat{\theta}_T - \theta| > h) \leq \alpha \tag{2.2}$$

where $\mathsf{P}(\cdot)$ and $\mathsf{E}[\cdot]$ denote probability and expectation including *averaging over* θ with the help of the prior.

Remark 2.1. We must emphasize that the constraint in (2.2) does not guarantee that the desired coverage probability will also hold for each individual θ, namely $\mathsf{P}(|\hat{\theta}_T - \theta| > h|\theta) \leq \alpha$, a property which is particularly desirable in practice. Perhaps, a more meaningful problem to consider in place of (2.2) would have been

$$\inf_{T,\hat{\theta}_T} \mathsf{E}[T], \text{ subject to: } \sup_{\theta} \mathsf{P}_\theta(|\hat{\theta}_T - \theta| > h) \leq \alpha, \tag{2.3}$$

that assures a coverage probability of at least $1-\alpha$ for *every* θ. Unfortunately, it is unclear how to derive the optimal solution to this alternative formulation. Consequently, we focus on (2.2) as the optimum scheme we are going to develop, but in our numerical examples, we will evaluate it in terms of (2.3) as well.

Let $c > 0$ denote a Lagrange multiplier that we use to combine the two terms in (2.2) into a single cost function $J(T,\hat{\theta}_T) = c\mathsf{E}[T] + \mathsf{P}(|\hat{\theta}_T - \theta| > h)$, and consider the *unconstrained* optimization problem

$$\inf_{T,\hat{\theta}_T} J(T,\hat{\theta}_T) = \inf_{T,\hat{\theta}_T} \left\{ c\mathsf{E}[T] + \mathsf{P}(|\hat{\theta}_T - \theta| > h) \right\}. \tag{2.4}$$

We will first identify the solution to (2.4) and then demonstrate that a proper selection of c can also solve the constrained problem in (2.2).

---
performance measure.

2.2.1 The Unconstrained Problem

We start by considering the classical Bayes estimation problem for fixed sample size t

$$\inf_{\hat{\theta}_t} \mathsf{P}(|\hat{\theta}_t - \theta| > h). \tag{2.5}$$

If we observe $\mathscr{F}_t = \sigma\{X_1, \ldots, X_t\}$ then, given that $\{X_t\}$ is i.i.d. Bernoulli(θ), the probability to obtain a specific combination of samples given θ is equal to $\theta^{S_t}(1-\theta)^{t-S_t}$, where $S_t = \sum_{k=1}^{t} X_k$ is the number of "successes" up to time t. This implies that the posterior probability density of θ given the observations can be written as

$$\pi_t(\theta|\mathscr{F}_t) = \pi_t(\theta|S_t) = \frac{\theta^{S_t}(1-\theta)^{t-S_t}\pi(\theta)}{\int_0^1 \theta^{S_t}(1-\theta)^{t-S_t}\pi(\theta)\,d\theta}. \tag{2.6}$$

From Bayesian estimation theory [25, Page 142], we have that the optimization in (2.5) is achieved by the following Bayes estimator

$$\hat{\vartheta}_t(S_t) = \arg\inf_{\hat{\theta}_t} \mathsf{P}(|\hat{\theta}_t - \theta| > h|\mathscr{F}_t) = \arg\sup_{\hat{\theta}_t} \int_{\max\{\hat{\theta}_t - h, 0\}}^{\min\{\hat{\theta}_t + h, 1\}} \pi_t(\theta|S_t)\,d\theta, \tag{2.7}$$

yielding the corresponding optimum conditional complementary coverage probability

$$\begin{aligned} \mathscr{C}_t(S_t) &= \inf_{\hat{\theta}_t} \mathsf{P}(|\hat{\theta}_t - \theta| > h|\mathscr{F}_t) = 1 - \sup_{\hat{\theta}_t} \int_{\max\{\hat{\theta}_t - h, 0\}}^{\min\{\hat{\theta}_t + h, 1\}} \pi_t(\theta|S_t)\,d\theta \\ &= 1 - \int_{\max\{\hat{\vartheta}_t(S_t) - h, 0\}}^{\min\{\hat{\vartheta}_t(S_t) + h, 1\}} \pi_t(\theta|S_t)\,d\theta. \end{aligned} \tag{2.8}$$

From (2.7) and (2.8) we observe that both quantities $\hat{\vartheta}_t(S_t), \mathscr{C}_t(S_t)$ are $\mathscr{F}_t$-measurable and, more precisely, functions of S_t. For known prior $\pi(\theta)$, we can, at least numerically, compute the Bayes estimate and the corresponding optimum conditional complementary coverage probability for each combination of (t, S_t).

Remark 2.2. By focusing on (2.7), we can make a small but interesting observation: Regarding the Bayes estimate $\hat{\vartheta}_t(S_t)$ it is easy to verify that

$$h \leq \hat{\vartheta}_t(S_t) \leq 1 - h. \tag{2.9}$$

Indeed, this is clear, because if in (2.7) we select $\hat{\theta}_t < h$ or $\hat{\theta}_t > 1 - h$, this will yield an inferior cost compared to the selection $\hat{\theta}_t = h$ or $\hat{\theta}_t = 1 - h$, respectively. The implication of this observation is that $\hat{\vartheta}_t(S_t)$ will be *biased and inconsistent* when considered as an estimate of the true parameter θ, at least for values of θ outside the interval $[h, 1 - h]$. As we mentioned, the correct meaning of this quantity is that it constitutes the *mid-point* of the confidence interval $[\hat{\vartheta}_t(S_t) - h, \hat{\vartheta}_t(S_t) + h]$ with the latter enjoying, for each fixed t, the largest possible coverage probability.

Consider now the optimization in (2.4) which will be performed in two steps: First we fix the stopping time T and minimize $J(T, \hat{\theta}_T)$ with respect to $\hat{\theta}_T$; the resulting expression is then minimized, during the second step, over T in order to obtain the optimum pair. We have the following lemma that addresses the first problem.

Lemma 2.1. *Assume stopping time T is fixed and satisfies $T \leq N$, where $N > 0$ is some deterministic integer. Then,*

$$J(T, \hat{\theta}_T) = c\mathsf{E}[T] + \mathsf{P}(|\hat{\theta}_T - \theta| > h) \geq \mathsf{E}[cT + \mathscr{C}_T] = \mathsf{J}(T), \tag{2.10}$$

with equality when we apply the corresponding Bayesian estimator $\hat{\theta}_T = \hat{\vartheta}_T$ at the time of stopping.

Proof. From (2.10) we can write

$$
\begin{aligned}
J(T,\hat{\theta}_T) = c\mathsf{E}[T] + \mathsf{P}(|\hat{\theta}_T-\theta|>h) &= \sum_{t=0}^{N}\mathsf{E}\left[\left\{ct+\mathbb{1}\{|\hat{\theta}_t-\theta|>h\}\right\}\mathbb{1}\{T=t\}\right] \\
&= \sum_{t=0}^{N}\mathsf{E}\left[\mathsf{E}\left[ct+\mathbb{1}\{|\hat{\theta}_t-\theta|>h\}|\mathscr{F}_t\right]\mathbb{1}\{T=t\}\right] && (2.11)\\
&= \sum_{t=0}^{N}\mathsf{E}\left[\left\{ct+\mathsf{P}(|\hat{\theta}_t-\theta|>h|\mathscr{F}_t)\right\}\mathbb{1}\{T=t\}\right] && (2.12)\\
&\geq \sum_{t=0}^{N}\mathsf{E}\left[\left\{ct+\inf_{\hat{\theta}_t}\mathsf{P}(|\hat{\theta}_t-\theta|>h|\mathscr{F}_t)\right\}\mathbb{1}\{T=t\}\right] && (2.13)\\
&= \sum_{t=0}^{N}\mathsf{E}\left[\{ct+\mathscr{C}_t\}\mathbb{1}\{T=t\}\right] && (2.14)\\
&= \mathsf{E}[cT+\mathscr{C}_T].
\end{aligned}
$$

The first equality in (2.12) is true because $\mathbb{1}\{T=t\}$ is $\mathscr{F}_t$-measurable, also we have equality in (2.14) if we select $\hat{\theta}_t=\hat{\vartheta}_t$ when $\{T=t\}$. We observe that changing the order of summation and expectation presents absolutely no complication because the stopping time is bounded by the deterministic quantity N. $\square$

A side-product of Lemma 2.1, as it can be verified from the corresponding proof, is the fact that the Bayesian estimator is not only optimum for fixed sample size, but it retains its optimality property when the sample size is controlled by any stopping time T adapted to the observations.

Using (2.10) from Lemma 2.1, we are now left with the optimization of the stopping time T. Assuming that N is an integer which is sufficiently large, we consider the following optimization over stopping times that are bounded by N

$$
\inf_{0\leq T\leq N}\mathsf{J}(T) = \inf_{0\leq T\leq N}\mathsf{E}[cT+\mathscr{C}_T]. \tag{2.15}
$$

This is a classical *finite horizon* optimal stopping problem with cost per sample equal to c

and cost for stopping at t equal to $\mathscr{C}_t$. Of course, it is only natural to wonder why we limited our analysis to finite horizons instead of considering the more classical *infinite horizon* version. As we will see in the sequel, for the most common prior we will be able to demonstrate that the infinite horizon assumption is completely unnecessary. Indeed, the optimum stopping time will turn out to be bounded by a deterministic quantity, suggesting that by limiting ourselves to a (sufficiently large) finite horizon, we do not suffer any performance loss.

In order to solve the optimization problem defined in (2.15), we follow the classical optimal stopping theory [26]. For $t = 0, 1, \ldots, N$ define the sequence of optimal *average residual costs*

$$\mathscr{V}_t = \inf_{t \leq T \leq N} \mathsf{E}[c(T-t) + \mathscr{C}_T | \mathscr{F}_t], \tag{2.16}$$

then we have

$$\mathscr{V}_t = \min\{\mathscr{C}_t, c + \mathsf{E}[\mathscr{V}_{t+1} | \mathscr{F}_t]\},\ t = N, \ldots, 1, 0, \tag{2.17}$$

with the backward recursion initialized with $\mathscr{V}_{N+1} = 1$. Regarding this last selection, it produces $\mathscr{V}_N = \mathscr{C}_N$ since the latter is a probability. In fact, this is exactly what the optimum residual cost at N must be, because if we have not stopped before N, then we necessarily stop at N and this produces cost $\mathscr{C}_N$ (simply the cost of stopping at N). The total optimum cost is expressed through $\mathscr{V}_0$, namely $\mathscr{V}_0 = \inf_{0 \leq T \leq N} \mathsf{J}(T)$. The next lemma specifies in more detail the recursion in (2.17).

Lemma 2.2. *Consider the recursion in* (2.17) *then, the optimal residual cost* $\mathscr{V}_t, t = N, \ldots, 0$ *is a function* $\mathscr{V}_t(S_t)$ *of* S_t *and therefore* $\mathscr{F}_t$*-measurable. Furthermore,* (2.17) *can be written as*

$$\mathscr{V}_t(S_t) = \min\{\mathscr{C}_t(S_t), c + \tilde{\mathscr{V}}_t(S_t)\},\ t = N, \ldots, 0, \tag{2.18}$$

where $\tilde{\mathscr{V}}_t(S_t)$ expresses the optimum average residual cost to continue, satisfying

$$\tilde{\mathscr{V}}_t(S_t) = g_{t+1}(S_t)\mathscr{V}_{t+1}(S_t+1) + \big(1 - g_{t+1}(S_t)\big)\mathscr{V}_{t+1}(S_t), \tag{2.19}$$

$$g_{t+1}(S_t) = \mathsf{P}(X_{t+1} = 1|\mathscr{F}_t) = \frac{\int_0^1 \theta^{S_t+1}(1-\theta)^{t-S_t}\pi(\theta)\,d\theta}{\int_0^1 \theta^{S_t}(1-\theta)^{t-S_t}\pi(\theta)\,d\theta}. \tag{2.20}$$

Finally, if the prior $\pi(\theta)$ is symmetric around $\frac{1}{2}$ then the functions $\mathscr{C}_t(S_t), \mathscr{V}_t(S_t), \tilde{\mathscr{V}}_t(S_t)$ are symmetric with respect to S_t around the value $\frac{t}{2}$.

Proof. The validity of this lemma is established in Section 2.6. □

Once the sequence of optimal residual costs has been obtained through the solution of (2.18), it is then immediate to define the optimum stopping time T_o that solves the minimization problem in (2.15). Again, optimal stopping theory [26] suggests that

$$T_\text{o} = \inf\{0 \le t \le N : \mathscr{V}_t(S_t) = \mathscr{C}_t(S_t)\} = \inf\{0 \le t \le N : \mathscr{C}_t(S_t) \le c + \tilde{\mathscr{V}}_t(S_t)\}. \tag{2.21}$$

In other words, when the optimum residual cost $\mathscr{V}_t(S_t)$ matches, for the first time, the cost for stopping $\mathscr{C}_t(S_t)$ or, equivalently, the cost of stopping is smaller than the residual cost of continuing, this is when we stop. Since the functions involved depend on S_t, this quantity can serve as our test statistic and we can express the stopping rule in (2.21) in terms of S_t. Specifically, for each time t, we can find the sampling region $\Omega_t = \{0 \le S_t \le t : \mathscr{V}_t(S_t) < \mathscr{C}_t(S_t)\} = \{0 \le S_t \le t : c + \tilde{\mathscr{V}}_t(S_t) < \mathscr{C}_t(S_t)\}$ with $\Omega_N = \varnothing$, and we can equivalently define the stopping time as $T_\text{o} = \inf\{0 \le t \le N : S_t \notin \Omega_t\}$.

2.2.2 The Constrained Problem

Let us now turn to the constrained problem in (2.2) which we can solve with the results we have so far. We will show that (2.2) can be recovered as an instance of the unconstrained version (2.4) corresponding to a special selection of the Lagrange multiplier c. Our result is summarized in the following theorem.

Theorem 2.1. *For the solution of* (2.2) *we distinguish two cases:*

i) If $\alpha \geq \mathscr{C}_0 = \mathsf{P}(|\hat{\vartheta}_0 - \theta| > h)$*, with* $\hat{\vartheta}_0 = \arg\inf_{\hat{\theta}_0} \mathsf{P}(|\hat{\theta}_0 - \theta| > h)$*, then the optimum is to stop without taking any samples, i.e.* $T_{\rm o} = 0$ *and use as mid-point of the optimum confidence interval the value* $\hat{\vartheta}_0$ *which is based only on the prior* $\pi(\theta)$*.*

ii) If $\mathsf{P}(|\hat{\vartheta}_0 - \theta| > h) > \alpha$*, then for any horizon* $N \geq N_\alpha$ *where* N_α *satisfies* $\mathsf{P}(|\hat{\vartheta}_{N_\alpha} - \theta| > h) < \alpha$*, there exists Lagrange multiplier* c_**, independent from* N*, such that the solution of* (2.4) *is also the solution to* (2.2) *that can involve a possible randomization before taking any samples.*

Proof. The proof of this theorem is presented in Section 2.6. □

2.3 Properties of the Optimum Solution

If we fix the value N of the horizon and the cost per sample c, we can then compute the mid-points $\{\{\hat{\vartheta}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ of the confidence intervals from (2.7). Assuming that $\pi(\theta)$ is continuous, candidates for $\hat{\vartheta}_t(S_t)$ can be obtained from the solution of the following equation which we obtain by differentiating (2.7) with respect to $\hat{\theta}_t$

$$(\hat{\theta}_t + h)^{S_t}(1 - \hat{\theta}_t - h)^{t-S_t}\pi(\hat{\theta}_t + h) - (\hat{\theta}_t - h)^{S_t}(1 - \hat{\theta}_t + h)^{t-S_t}\pi(\hat{\theta}_t - h) = 0. \quad (2.22)$$

The previous equation has clearly a solution in the interval $[h, 1-h]$ when $0 < S_t < t$ with the corresponding value providing a (local) extremum for the coverage probability. To these candidate mid-points we must include the two end points $h, 1-h$ since the global maximum can occur at the two ends as well. Therefore, we need to examine which of these cases provides the best coverage probability and select the corresponding value as our optimum mid-point $\hat{\vartheta}_t(S_t)$. When $S_t = 0, t$ it is possible (2.22) not to have any solution in $[h, 1-h]$. In this case, $\hat{\vartheta}_t(0)$ and $\hat{\vartheta}_t(t)$ are equal to one of the two end values h or $1-h$. Having identified the optimum mid-points $\{\{\hat{\vartheta}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$, we apply (2.8) to compute the corresponding optimum complementary conditional coverage probabilities

$\{\{\mathscr{C}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$.

The next step consists in computing $\{\{g_{t+1}(S_t)\}_{S_t=0}^t\}_{t=0}^N$ for $t = 0, \ldots, N$ and $S_t = 0, \ldots, t$ with numerical integration. Once we have the available sequences $\{\{\mathscr{C}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ and $\{\{g_{t+1}(S_t)\}_{S_t=0}^t\}_{t=0}^N$, we can then use them in the backward recursion (2.18) to find the sequence $\{\{\tilde{\mathscr{V}}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ and the optimum residual cost sequence $\{\{\mathscr{V}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$. To identify the stopping rule, according to (2.21) we must compare the two sequences $\{\{\mathscr{C}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$, $\{\{\mathscr{V}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ element-by-element. At coordinates (t, S_t) where the sequences differ, we decide to continue sampling; whereas if they are equal, we decide to stop. This generates the sequence of sampling regions $\{\Omega_t\}_{t=0}^N$. Equivalently, we can compare $\{\{\mathscr{C}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ with $\{\{c + \tilde{\mathscr{V}}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$, and wherever the first is no larger than the second, we stop, while we continue sampling in the opposite case.

We now present a conjecture that contains two significant claims for the optimum stopping time for the problem in (2.4) which we believe are valid for *any* prior $\pi(\theta)$. We were able to provide a proof for the first claim (Lemma 2.3) for a rich class of priors, and prove both claims (Theorem 2.2) providing also quantitative information when the prior is the Beta density. Regarding the latter case we should note that the Beta density is among the most popular priors for the problem we are considering in this work.

Conjecture. *For any prior $\pi(\theta)$ and sufficiently large horizon N the optimum stopping time $T_{\rm o}$ of the unconstrained problem in* (2.4) *enjoys the following two properties:*
i). There exists constant $t_{\rm up}$ depending only on c and not on N such that $T_{\rm o} \leq t_{\rm up}$.
ii). For sufficiently small c there exists constant $t_{\rm lo} \geq 1$ depending only on c and not on N such that $t_{\rm lo} \leq T_{\rm o}$.

Below we present a general proof of property i) of the Conjecture under the following additional assumption: Define the maximal conditional variance

$$\sigma_t^2 = \max_{0 \leq S_t \leq t} \mathsf{E}\left[\left(\theta - \mathsf{E}[\theta|S_t]\right)^2|S_t\right] = \max_{0 \leq S_t \leq t} \int_0^1 (\theta - \mathsf{E}[\theta|S_t])^2 \pi_t(\theta|S_t) d\theta, \tag{2.23}$$

where $\pi_t(\theta|S_t)$ is the posterior pdf defined in (2.6) and assume that $\sigma_t \to 0$ as $t \to \infty$. This forces the conditional variance to converge to 0 *uniformly* in S_t. It also implies that the posterior distribution $\pi_t(\theta|S_t)$ converges, uniformly, to a degenerate measure at a single point (often the true θ) as $t \to \infty$. This is clearly related to the *consistency* concept of posterior distributions in Bayesian statistics and is often considered a valid assumption (see [27]).

Lemma 2.3. *Let σ_t be defined as in (2.23) with $\lim_{t\to\infty} \sigma_t = 0$. Then for sufficiently large horizon there exists constant t_{up} depending only on c such that $T_{\text{o}} \leq t_{\text{up}}$, i.e. property i) in the Conjecture is true.*

Proof. The proof is a simple application of the Chebyshev inequality in combination with (2.23). Indeed we observe that

$$\mathscr{C}_t(S_t) = \inf_{\hat{\theta}_t} \mathsf{P}(|\theta - \hat{\theta}_t| > h|\mathscr{F}_t) \leq \mathsf{P}\big(|\theta - \mathsf{E}[\theta|\mathscr{F}_t]| > h|\mathscr{F}_t\big)$$
$$\leq \frac{1}{h^2}\mathsf{E}\big[\big(\theta - \mathsf{E}[\theta|S_t]\big)^2|S_t\big] \leq \frac{\sigma_t^2}{h^2}. \quad (2.24)$$

Since $\sigma_t \to 0$ as $t \to \infty$, there exists N such that $\mathscr{C}_N \leq \frac{\sigma_N^2}{h^2} \leq c$ and, therefore, from (2.18) we conclude that $\mathscr{C}_N \leq c + \tilde{\mathscr{V}}_N$, which suggests that we will necessarily stop at N for any value of S_N. Quantity t_{up} is the smallest N for which this is true. $\square$

Remark 2.3. The assumption $\lim_{t\to\infty} \sigma_t = 0$ does not hold for all prior distribution. A counterexample where it fails is when the prior is a two-point probability mass function, say $\mathsf{P}(\theta = 0.4) = \mathsf{P}(\theta = 0.6) = 0.5$. However, even for this case the Conjecture might still be valid since the requirement $\mathscr{C}_N(S_N) \leq \frac{\sigma_N^2}{h^2} < c$ used in our proof, is only sufficient for the validity of our claim.

An interesting example where the assumption holds is when the prior is the Beta density

$\pi(\theta) = \text{Beta}(\theta, p, q)$, where

$$\text{Beta}(\theta, p, q) = \frac{\theta^{p-1}(1-\theta)^{q-1}}{\int_0^1 \theta^{p-1}(1-\theta)^{q-1}\,d\theta}, \quad p, q > 0. \tag{2.25}$$

To see this, we note that the posterior pdf is of the same type, namely $\pi(\theta|S_t) = \text{Beta}(\theta, p+S_t, t-S_t+q)$, and thus the maximal conditional variance in (2.23) becomes

$$\sigma_t^2 = \max_{0\le S_t\le t} \frac{(p+S_t)(t-S_t+q)}{(t+p+q)^2(t+p+q+1)} \le \frac{1}{4(t+p+q+1)}, \tag{2.26}$$

where the equality is attainable when $S_t = \frac{t+q-p}{2}$ is an integer. Clearly, for fixed $p, q > 0$ we have $\sigma_t \to 0$ as $t \to \infty$, and thus the assumption of Lemma 2.3 holds. Moreover, by the proof of Lemma 2.3, the optimum stopping time satisfies $T_\text{o} \le \max\{0, \frac{1}{4h^2c} - p - q - 1\}$ for all $c > 0$. This bound is of the order of c^{-1}. In Theorem 2.2, Section 2.3.2, by applying a more advanced analysis, we will be able to improve it and provide an alternative estimate which is of the order of $|\log(c)|$ for the case of the symmetric prior $p = q$.

Remark 2.4. Property i) of the Conjecture suggests that the number of samples, under the optimum scheme, will never exceed the value t_up even if we allow the horizon to grow without limit. This interesting and uncommon characteristic was also observed in [24] but with cost function a variance of the classical mean square error. However, what is more intriguing in our conjecture is property ii), namely that we need first to accumulate a sufficient volume of information before we start asking ourselves whether we should stop sampling or not. This is an extremely uncommon feature and, to our knowledge, has never been reported before in Sequential Analysis as a property of optimum schemes. As we claim in our conjecture, we believe that both properties are valid for any prior $\pi(\theta)$. Fortunately, as we mentioned before, this double claim is not without solid evidence. Indeed with Theorem 2.2 (Page 23), we demonstrate its validity when the prior is the symmetric Beta density.

2.3.1 Performance Evaluation

What we presented so far allows for the determination of the stopping rule of the proposed scheme. We would like now to compute its performance but also the performance of any stopping time which uses S_t as its test statistic and is defined in terms of a sequence of sampling regions $\{\Omega_t\}$ in terms of $\{S_t\}$. In particular, we are interested in computing $\mathsf{E}_\theta[T], \mathsf{E}[T], \mathsf{P}_\theta(|\hat{\theta}_T - \theta| \leq h)$ and $\mathsf{P}(|\hat{\theta}_T - \theta| \leq h)$. Of course, we could obtain these quantities using Monte-Carlo simulations, but it is also possible to determine them numerically. The following lemma provides the necessary formulas.

Lemma 2.4. *Let the stopping time T be bounded by N having as test statistic the process $\{S_t\}$. Assume for each t that Ω_t denotes the sampling region. Suppose also that for the combination (t, S_t) the scheme provides the mid-point estimate $\hat{\theta}_t(S_t)$ and the corresponding conditional complementary coverage probability $C_t(S_t) = \mathsf{P}(|\hat{\theta}_t(S_t) - \theta| > h|\mathscr{F}_t)$. For $t = N-1, \ldots, 0$, we then define the following backward recursions that must be applied for $S_t = 0, 1, \ldots, t$*

$$U_t(S_t) = 1 + \theta \mathbb{1}\{S_t + 1 \in \Omega_{t+1}\} U_{t+1}(S_t + 1) + (1-\theta)\mathbb{1}\{S_t \in \Omega_{t+1}\} U_{t+1}(S_t), \tag{2.27}$$

$$\begin{aligned} \bar{U}_t(S_t) = 1 &+ g_{t+1}(S_t)\mathbb{1}\{S_t + 1 \in \Omega_{t+1}\}\bar{U}_{t+1}(S_t+1) \\ &+ \big(1 - g_{t+1}(S_t)\big)\mathbb{1}\{S_t \in \Omega_{t+1}\}\bar{U}_{t+1}(S_t), \end{aligned} \tag{2.28}$$

$$\begin{aligned} W_t(S_t) = &\ \mathbb{1}\{|\hat{\theta}_t - \theta| > h\}\mathbb{1}\{S_t \notin \Omega_t\} \\ &+ \big\{\theta W_{t+1}(S_t+1) + (1-\theta)W_{t+1}(S_t)\big\}\mathbb{1}\{S_t \in \Omega_t\}, \end{aligned} \tag{2.29}$$

$$\begin{aligned} \bar{W}_t(S_t) = &\ C_t(S_t)\mathbb{1}\{S_t \notin \Omega_t\} + \big\{g_{t+1}(S_t)\bar{W}_{t+1}(S_t+1) \\ &+ \big(1 - g_{t+1}(S_t)\big)\bar{W}_{t+1}(S_t)\big\}\mathbb{1}\{S_t \in \Omega_t\}, \end{aligned} \tag{2.30}$$

where $g_{t+1}(S_t)$ is defined in (2.20) and the four recursions are initialized with $U_N(S_N) = \bar{U}_N(S_N) = 0, W_N(S_N) = \mathbb{1}\{|\hat{\theta}_N - \theta| > h\}, \bar{W}_N(S_N) = C_N(S_N), \Omega_N = \varnothing$. Then, $\mathsf{E}_\theta[T] = U_0(S_0), \mathsf{E}[T] = \bar{U}_0(S_0), \mathsf{P}_\theta(|\hat{\theta}_T - \theta| > h|) = W_0(S_0)$ and $\mathsf{P}(|\hat{\theta}_T - \theta| > h) =$

$\bar{W}_0(S_0)$.

Proof. The validity of these expressions is established in Section 2.6. □

The applicability of Lemma 2.4 is clearly not limited to the proposed scheme but can be used to compute the performance of the fixed-sample-size and of other sequential alternatives that we intend to compare against the method we have developed.

2.3.2 Beta Density as Prior

Let us now find the particular form of our scheme when we adopt as our prior the Beta density $\pi(\theta) = \text{Beta}(\theta, a, a)$, where $\text{Beta}(\theta, p, q)$ is defined in (2.25). We observe that the selection $a = 1$ in the prior corresponds to the uniform density in $[0, 1]$. It is now straightforward to verify that the posterior pdf accepts a similar form, namely

$$\pi(\theta|S_t) = \text{Beta}(\theta, a + S_t, a + t - S_t), \tag{2.31}$$

while the conditional complementary coverage probability at time t becomes

$$\mathsf{P}(|\hat{\theta}_t - \theta| > h|\mathscr{F}_t) = 1 - I_{\min\{1,\hat{\theta}_t+h\}}(a + S_t, a + t - S_t) + I_{\max(0,\hat{\theta}_t-h)}(a + S_t, a + t - S_t), \tag{2.32}$$

where $I_x(p, q)$ is the incomplete Beta function (see [28, Page 944]) which is the cdf of $\text{Beta}(\theta, p, q)$.

The Bayes estimator, according to (2.22), can be found as the solution of the equation

$$\hat{\vartheta}_t = \arg\left\{\hat{\theta}_t : \left(\frac{\hat{\theta}_t - h}{\hat{\theta}_t + h}\right)^{a+S_t-1} = \left(\frac{1 - h - \hat{\theta}_t}{1 + h - \hat{\theta}_t}\right)^{a+t-S_t-1}\right\}$$

corresponding to the root in the interval $[h, 1 - h]$. Such root always exists except when $a = 1$ and $S_t = 0$ or t. For these cases, $\hat{\vartheta}_t$ is equal to h or $1 - h$, depending on which value provides a larger conditional coverage probability. The resulting optimum conditional complementary coverage probability becomes

$$\mathscr{C}_t(S_t) = 1 - I_{\min\{1,\hat{\vartheta}_t+h\}}(a+S_t, a+t-S_t) + I_{\max(0,\hat{\vartheta}_t-h)}(a+S_t, a+t-S_t). \quad (2.33)$$

Finally, as indicated in (2.19) and (2.20), we need to find the probability $g_{t+1}(S_t)$, for which we have the following simple formula

$$g_{t+1}(S_t) = \mathsf{P}(X_{t+1}=1|\mathscr{F}_t) = \frac{\Gamma(S_t+a+1)\Gamma(t+2a)}{\Gamma(S_t+a)\Gamma(t+2a+1)} = \frac{S_t+a}{t+2a}. \quad (2.34)$$

We can now compute the sequences $\{\{\mathscr{V}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N, \{\{\tilde{\mathscr{V}}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ as explained in (2.18) and compare, element-by-element, $\{\{\mathscr{C}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ with $\{\{\mathscr{V}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ or $\{\{\mathscr{C}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ with $\{\{c+\tilde{\mathscr{V}}_t(S_t)\}_{S_t=0}^t\}_{t=0}^N$ to identify the sampling and stopping regions.

For the particular prior adopted in (2.25), as we mentioned before, the resulting optimum stopping time T_o enjoys the unique properties claimed in the Conjecture. The next theorem provides the necessary evidence.

Theorem 2.2. *The Conjecture is true when the prior is the Beta density $\pi(\theta) =$ Beta(θ, a, a) with the optimum stopping time T_o satisfying $C_0|\log(c)| \le T_\mathrm{o} \le C_1|\log(c)|$ for constants $C_0 < C_1$ that depend only on a and h.*

Proof. The proof is very technical and detailed in Section 2.6. Unfortunately, the analytical techniques developed for the specific prior are not directly extendable to the general case. □

Perhaps, it is worth mentioning the fact that from the proof of Theorem 2.2, we conclude that the two estimates for t_up and t_lo in (2.51),(2.53) grow linearly in $|\log(c)|$ having drastically different multiplicative coefficients (C_0 of the order of $\frac{1}{|\log(0.5-h)|}$ versus C_1 of the order of $\frac{1}{2h^2}$) and different offsets.

In Fig. 2.2, after using (2.28), we plot the average sample size and the two limits $t_\mathrm{lo}, t_\mathrm{up}$

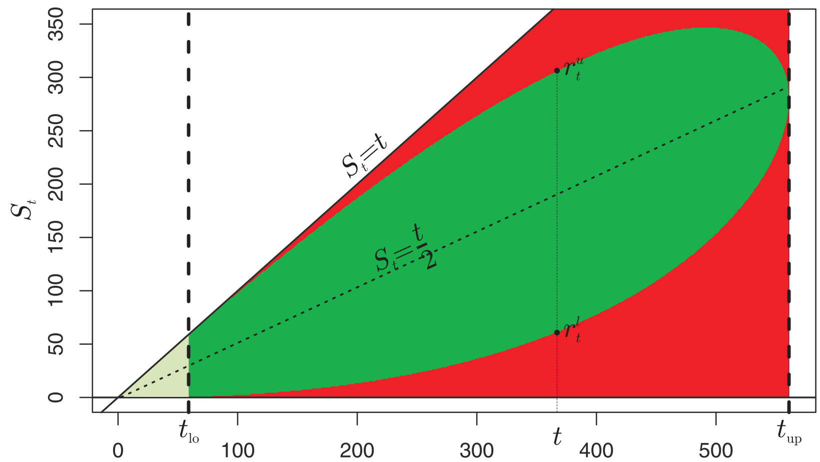


Figure 2.1: Sampling (green) and stopping (red) regions for $a = 1, h = 0.05$ and $c = 0.0001$. Upper and lower bounds for optimum stopping time: $t_{\text{lo}} = 59$ and $t_{\text{up}} = 561$. No possibility of stopping (light green).

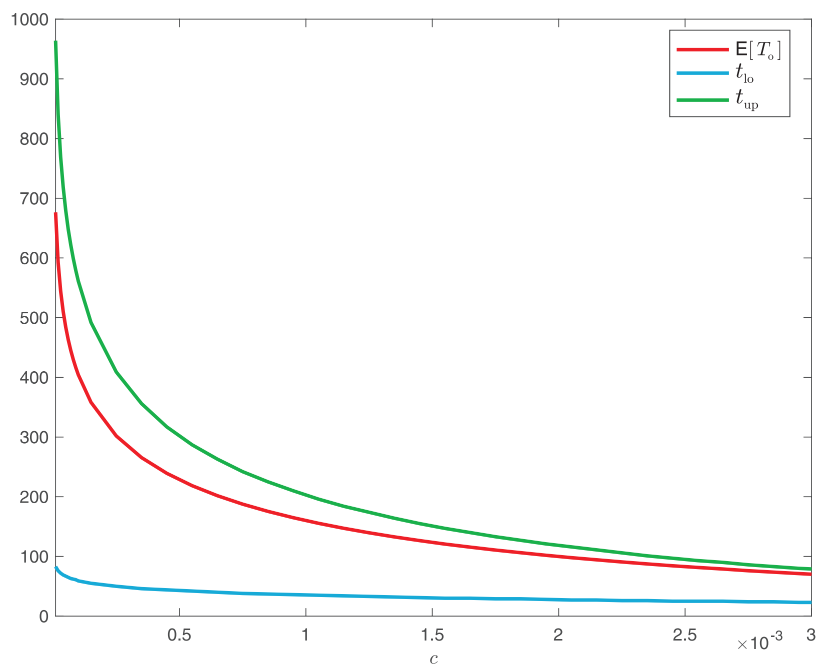


Figure 2.2: Average sample size (red), lower t_{lo} (blue) and upper t_{up} limit (green), as functions of c for optimum stopping time T_{o} when $a = 1$ and $h = 0.05$.

of T_o as functions of c for $a = 1$ and $h = 0.05$. We can see that the lower limit t_lo is significantly smaller than the resulting average, suggesting that the optimum scheme very quickly regards the accumulated information as capable of providing reliable interval estimates and therefore starts the process of questioning whether to stop or continue sampling.

As an illustration for these properties we consider $a = 1, h = 0.05$, and $c = 0.0001$. Fig. 2.1 depicts the sampling (green) and the stopping (red) region in terms of the test statistic S_t. Both regions are clearly limited between the lines $S_t = t$ and $S_t = 0$. Even though we have marked a whole region in red, only the points that are next to the green region are actually accessible because S_t can increase at most by one unit as we go from t to $t + 1$. We can also see the two bounds $t_\mathrm{up} = 561$ and $t_\mathrm{lo} = 59$ for T_o. For $t \le t_\mathrm{lo}$ the light green region covers all points $0 \le S_t \le t$, thus identifying the time instances we can never stop. Also, we note that once we pass t_up we are in the stopping region suggesting that we must necessarily stop at t_up. For each $t_\mathrm{lo} \le t \le t_\mathrm{up}$ the stopping region has an upper r_t^u and a lower r_t^l threshold and, as long as S_t is between these two limits, we need to sample. Since the prior distribution is symmetric with respect to $1/2$, then, according to Theorem 2.1, the sampling region is symmetric around $t/2$, implying that $r_t^u + r_t^l = t$.

2.4 Comparisons

Let us now compare our scheme with the optimal fixed-sample-size (FSS) and two sequential methods: The first was proposed by Frey in [23] and the second, the *Conditional Method*, was proposed in our earlier work in [29] and in Chapter 4. For more details about this method, please refer to Chapter 4 in this thesis. Frey's method uses a modified Wald-type sequential confidence interval based on the stopping time

$$T_\mathrm{F} = \inf\left\{ t \ge 0 : \frac{\tilde{\theta}_{t,k}(1-\tilde{\theta}_{t,k})}{t} \le \left(\frac{h}{z_{\frac{\gamma}{2}}}\right)^2 \right\}, \tag{2.35}$$

where $\tilde{\theta}_{t,k} = \frac{S_t+k}{t+2k}$, $k > 0$ is a pre-specified constant and $\gamma = \gamma(k, h, \alpha)$ is chosen so that the confidence interval $[\hat{\theta}_{T_{\mathrm{F}}} - h, \hat{\theta}_{T_{\mathrm{F}}} + h]$, with $\hat{\theta}_t = \frac{S_t}{t}$, has a confidence level of at least $1 - \alpha$. Table 2.1 provides the values of k and γ recommended in [23] for best results. From (2.35) and using the fact that $x(1 - x) \leq \frac{1}{4}$ we conclude that the corresponding

Table 2.1: Choices of k and γ for 90%, 95%, and 99% confidence intervals of fixed half-width h in [23].

	90%		95%		99%	
h	k	γ	k	γ	k	γ
0.10	4	0.0754	4	0.0356	6	0.0068
0.05	4	0.0859	6	0.0433	8	0.0083
0.01	8	0.0972	10	0.0487	14	0.0097

stopping time satisfies $T_{\mathrm{F}} \leq \lceil \frac{z_{\frac{\gamma}{2}}^2}{4h^2} \rceil = N$. Regarding the fixed-sample-size method, it uses the optimum Bayes estimator $\hat{\vartheta}_t$, obtained in (2.7) and the number of samples t is selected to meet the desired coverage probability. Finally, for the conditional method in [29] and in Chapter 4, we should point out that it is a general sequential parameter estimation technique based on conditional costs which is not limited to binomial proportions. For the problem of interest, we have $T_{\mathrm{C}} = \inf\{t \geq 0 : \mathscr{C}_t \leq \beta\}$ and $\hat{\theta}_{T_{\mathrm{C}}} = \hat{\vartheta}_{T_{\mathrm{C}}}$, where $\hat{\vartheta}_t, \mathscr{C}_t$ are the Bayes estimator and the corresponding optimum conditional complementary coverage probability defined in (2.7) and (2.8). Threshold β is selected to guarantee that the resulting coverage probability is $1 - \alpha$. For $\mathscr{C}_t$ we have from the proof of Theorem 2.1, eq. (2.12), that $\mathscr{C}_t \leq 2e^{-2h^2(t+2a+1)}$, consequently $T_{\mathrm{C}} \leq \lceil \max\{\frac{|\log(\frac{\beta}{2})|}{2h^2} - 2a - 1, 0\} \rceil = N$. In other words, all four schemes satisfy the assumption of Lemma 2.4 of bounded stopping time, therefore the corresponding performance can be computed numerically by applying the recursions of the lemma without the need to perform Monte-Carlo simulations.

For the competing methods using (2.28) and (2.30), we plot in Fig. 2.3 the average number of samples $\mathsf{E}[T]$ versus the coverage probability $\mathsf{P}(|\hat{\theta}_T - \theta| \leq h)$ when $a = 1$ and $h = 0.05$. Note that we have three points for Frey's scheme because of the tuning parameters k and γ which are provided in Table 2.1 only for three confidence levels. As

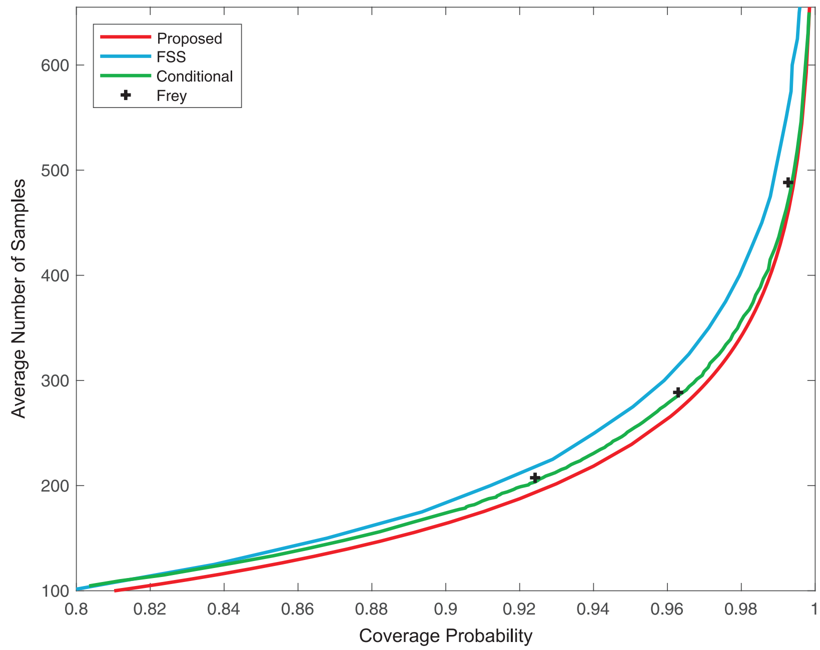
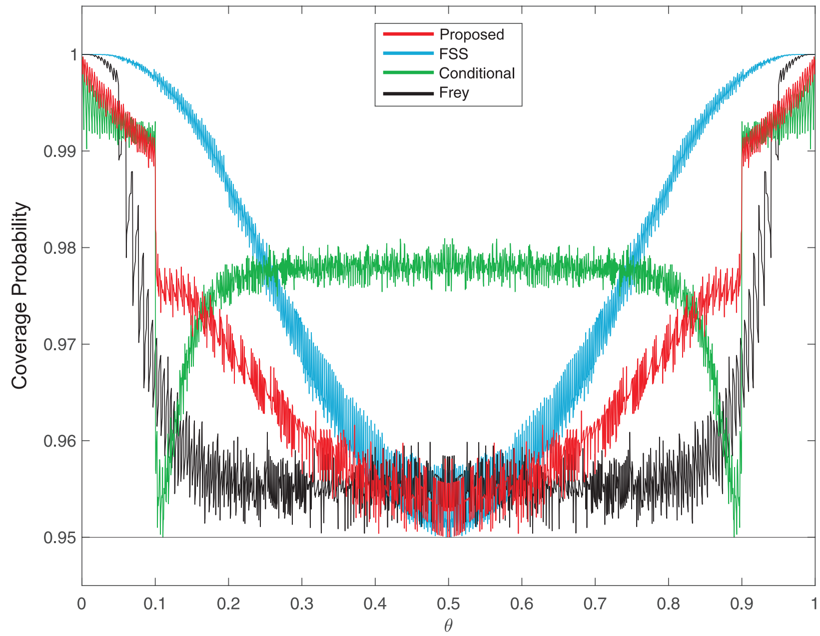


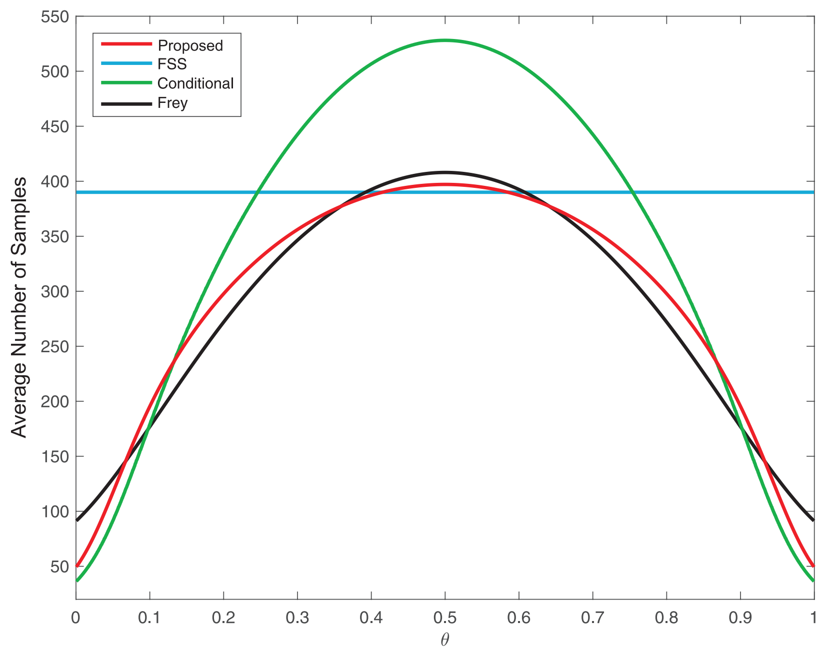
Figure 2.3: Average samples size versus coverage probability for proposed (red), Frey (black $+$), fixed-sample-size (blue) and conditional (green), for $a = 1$ and $h = 0.05$.

we can see, the proposed method outperforms the fixed-sample-size and both alternative sequential techniques. It is only at very high coverage probability levels that the difference between the three sequential schemes becomes less pronounced.

As we pointed out in (2.3), Section 2.2, there is practical interest in evaluating the performance for each individual θ. Clearly in this case, the requirement is to be able to guarantee a minimal coverage probability *for all* θ. Again, we resort to Lemma 2.4 and use (2.27) and (2.29) to evaluate the performance of the competing methods for each θ. In Fig. 2.4a, we plot the coverage probability for each test versus θ and in Fig. 2.4b, the corresponding average sample size required to obtain this performance. Parameters were selected so as all competing schemes provide *the same worst-case coverage probability* assuring a coverage of at least 0.95 for all θ. By observing the two figures, we can draw the following conclusions: The fixed-sample-size scheme can require up to almost eight times more samples compared to the proposed. Of course, one may argue that it produces higher coverage probability levels. Indeed this is true, but, unfortunately, this increased performance cannot be traded for a reduced sample size without compromising the worst-case level. Conse-



(a)



(b)

Figure 2.4: Coverage probability (a) and Average sample size (b) as a function of proportion θ for proposed (red), Frey (black), fixed-sample-size (blue) and conditional (green) when $a = 1$, $h = 0.05$ and worst-case coverage probability 0.95.

quently, what we observe is in fact the best the fixed-sample-size method can offer. The conditional scheme, around $\theta = 0.5$, requires up to 30% more samples which, as in the case of fixed-sample-size, produce higher coverage probabilities. Again, it is impossible to sacrifice part of this increased performance to improve the corresponding sample size without degrading the worst-case coverage probability. Finally, we can see that the proposed and Frey's scheme require similar samples over most θ. However, we observe that the proposed method has a coverage probability profile which is better than Frey's, since for most θ the corresponding probability is larger. Frey's scheme is slightly better only for θ close to 0 and 1. But even for these values of θ the proposed scheme requires almost 50% less samples.

2.5 Conclusions

We proposed an optimal sequential scheme for obtaining confidence intervals for a binomial proportion under a well defined formulation. We proved that, for a particular prior (Beta density), our optimum stopping time enjoys certain uncommon properties not encountered in solutions of other classical optimal stopping problems. We also conjectured that these properties are present with any prior. Specifically, our claim is that our stopping time is always bounded from above and below, suggesting that we need to first accumulate a sufficient amount of information before we start applying our stopping rule, and that our stopping time will always terminate before or at a specific deterministic time even if we allow the time horizon to be infinite. Finally, our scheme was compared against the optimum fixed-sample-size procedure and against existing sequential alternatives. Numerical performance evaluations showed that the proposed method exhibits an overall improved performance profile compared to its rivals.

2.6 Chapter 2 Proofs

Proof of Lemma 2.2: Let us start first with the recursion in (2.17). To simplify relation (2.17), it is crucial to take advantage of the relationship between S_t and S_{t+1} : $S_{t+1} = S_t + 1$ if $X_{t+1} = 1$ and $S_{t+1} = S_t$ if $X_{t+1} = 0$. Thus, the expectation term in (2.17) actually reduces to the conditional probability mass function of the Bernoulli variable X_{t+1} conditional on $\mathscr{F}_t$ when θ has a prior pdf $\pi(\theta)$. Denote

$$g_{t+1}(S_t) = \mathsf{P}(X_{t+1} = 1|\mathscr{F}_t) = \mathsf{P}(X_{t+1} = 1|S_t)$$

as S_t is a sufficient statistic for θ. Hence,

$$\begin{aligned} g_{t+1}(S_t) = \mathsf{P}(X_{t+1} = 1|\mathscr{F}_t) &= \frac{\int_0^1 \mathsf{f}(x_1, \ldots, x_t, X_{t+1} = 1|\theta)\pi(\theta)d\theta}{\int_0^1 \mathsf{f}(x_1, \ldots, x_t|\theta)\pi(\theta)d\theta} \\ &= \frac{\int_0^1 \theta^{S_t+1}(1-\theta)^{t-S_t}\pi(\theta)d\theta}{\int_0^1 \theta^{S_t}(1-\theta)^{t-S_t}\pi(\theta)d\theta}. \end{aligned} \tag{2.36}$$

Furthermore, $\mathsf{P}(X_{t+1} = 0|\mathscr{F}_t) = \mathsf{P}(X_{t+1} = 0|S_t) = 1 - g_{t+1}(S_t)$, and relation (2.17) becomes

$$\mathscr{V}_t(S_t) = \min\{\mathscr{C}_t(S_t), c + \mathscr{V}_{t+1}(S_t+1)g_{t+1}(S_t) + \mathscr{V}_{t+1}(S_t)(1 - g_{t+1}(S_t))\}, \tag{2.37}$$

for $t = N, N-1, \ldots, 0$.

To prove that functions $\mathscr{C}_t(S_t), \mathscr{V}_t(S_t), \tilde{\mathscr{V}}_t(S_t)$ are symmetric with respect to S_t around the value $t/2$, let $U_t = 2S_t - t$. One can think of $U_t = S_t - (t - S_t)$ as the number of differences between 0's and 1's in the first t trials. Then, suffices to show that $\mathscr{C}_t(U_t) = \mathscr{C}_t(-U_t)$ and $\mathscr{V}(U_t) = \mathscr{V}(-U_t)$ for all $t > 0$.

First, if $\pi(\theta)$ is symmetric around $1/2$, then

$$\pi(\theta) = \pi(1-\theta), \text{ for all } \theta \in (0,1). \tag{2.38}$$

Then, using (2.6), the posterior probability density of θ can be written as

$$\pi_t(\theta|U_t) = \frac{\theta^{U_t/2+t/2}(1-\theta)^{t/2-U_t/2}\pi(\theta)}{\int_0^1 \theta^{U_t/2+t/2}(1-\theta)^{t/2-U_t/2}\pi(\theta)\,d\theta}.$$

Applying change of measure from θ to $1-\theta$ and using (2.38) yield

$$\begin{aligned}
\pi_t(1-\theta|-U_t) &= \frac{(1-\theta)^{-U_t/2+t/2}\theta^{t/2+U_t/2}\pi(1-\theta)}{\int_0^1 (1-\theta)^{-U_t/2+t/2}\theta^{t/2+U_t/2}\pi(1-\theta)\,d\theta} \\
&= \frac{\theta^{U_t/2+t/2}(1-\theta)^{t/2-U_t/2}\pi(\theta)}{\int_0^1 \theta^{U_t/2+t/2}(1-\theta)^{t/2-U_t/2}\pi(\theta)\,d\theta} \\
&= \pi_t(\theta|U_t). \qquad (2.39)
\end{aligned}$$

Furthermore, using (2.7) and (2.39), we have

$$\begin{aligned}
1-\hat{\vartheta}_t(-U_t) &= \arg\sup_{\hat{\theta}_t} \int_{\max\{1-\hat{\theta}_t-h,0\}}^{\min\{1-\hat{\theta}_t+h,1\}} \pi_t(\theta|-U_t)\,d\theta \\
&= \arg\sup_{\hat{\theta}_t} \int_{1-\max\{1-\hat{\theta}_t-h,0\}}^{1-\min\{1-\hat{\theta}_t+h,1\}} -\pi_t(1-\theta|-U_t)\,d\theta \\
&= \arg\sup_{\hat{\theta}_t} \int_{\max\{\hat{\theta}_t-h,0\}}^{\min\{\hat{\theta}_t+h,1\}} \pi_t(\theta|U_t)\,d\theta \\
&= \hat{\vartheta}_t(U_t). \qquad (2.40)
\end{aligned}$$

Consequently, using (2.8) and (2.40),

$$\begin{aligned}
\mathscr{C}_t(-U_t) &= 1-\int_{\max\{\hat{\vartheta}_t(-U_t)-h,0\}}^{\min\{\hat{\vartheta}_t(-U_t)+h,1\}} \pi_t(\theta|-U_t)\,d\theta \\
&= 1-\int_{1-\max\{\hat{\vartheta}_t(-U_t)-h,0\}}^{1-\min\{\hat{\vartheta}_t(-U_t)+h,1\}} -\pi_t(1-\theta|-U_t)\,d\theta \\
&= 1-\int_{\max\{\hat{\vartheta}_t(U_t)-h,0\}}^{\min\{\hat{\vartheta}_t(U_t)+h,1\}} \pi_t(\theta|U_t)\,d\theta \\
&= \mathscr{C}_t(U_t). \qquad (2.41)
\end{aligned}$$

Now, using (2.36) and (2.38),

$$\begin{aligned}
&g_{t+1}(U_t) + g_{t+1}(-U_t) \\
&= \frac{\int_0^1 \theta^{U_t/2+t/2+1}(1-\theta)^{t/2-U_t/2}\pi(\theta)d\theta}{\int_0^1 \theta^{U_t/2+t/2}(1-\theta)^{t/2-U_t/2}\pi(\theta)d\theta} + \frac{\int_0^1 \theta^{-U_t/2+t/2+1}(1-\theta)^{t/2+U_t/2}\pi(\theta)d\theta}{\int_0^1 \theta^{-U_t/2+t/2}(1-\theta)^{t/2+U_t/2}\pi(\theta)d\theta} \\
&= \frac{\int_0^1 \theta^{U_t/2+t/2+1}(1-\theta)^{t/2-U_t/2}\pi(\theta)d\theta}{\int_0^1 \theta^{U_t/2+t/2}(1-\theta)^{t/2-U_t/2}\pi(\theta)d\theta} + \frac{\int_0^1 (1-\theta)^{-U_t/2+t/2+1}\theta^{t/2+U_t/2}\pi(1-\theta)d\theta}{\int_0^1 (1-\theta)^{-U_t/2+t/2}\theta^{t/2+U_t/2}\pi(1-\theta)d\theta} \\
&= \frac{\int_0^1 \theta^{U_t/2+t/2}(1-\theta)^{t/2-U_t/2}(\theta+1-\theta)\pi(\theta)d\theta}{\int_0^1 \theta^{U_t/2+t/2}(1-\theta)^{t/2-U_t/2}\pi(\theta)d\theta} = 1.
\end{aligned} \tag{2.42}$$

Finally, at time N, we have

$$\mathscr{V}_N(-U_N) = \mathscr{C}_N(-U_N) = \mathscr{C}_N(U_N) = \mathscr{V}_N(U_N).$$

Assume that for some $t = N-1, N-2, \ldots, 0$, $\mathscr{V}_{t+1}(U_{t+1}) = \mathscr{V}_{t+1}(-U_{t+1})$. Then, using the fact that $U_{t+1} = 2(S_t + X_{t+1}) - (t+1) = U_t + 2X_{t+1} - 1$ and (2.37),

$$\mathscr{V}_t(U_t) = \min\{\mathscr{C}_t(U_t), c + \mathscr{V}_{t+1}(U_t+1)g_{t+1}(U_t) + \mathscr{V}_{t+1}(U_t-1)(1-g_{t+1}(U_t))\}.$$

Moreover,

$$\begin{aligned}
\mathscr{V}_t(-U_t) &= \min\{\mathscr{C}_t(-U_t), c + \mathscr{V}_{t+1}(-U_t+1)g_{t+1}(-U_t) + \mathscr{V}_{t+1}(-U_t-1)(1-g_{t+1}(-U_t))\} \\
&= \min\{\mathscr{C}_t(U_t), c + \mathscr{V}_{t+1}(U_t-1)(1-g_{t+1}(U_t)) + \mathscr{V}_{t+1}(U_t+1)g_{t+1}(U_t)\} \\
&= \mathscr{V}_t(U_t).
\end{aligned}$$

This concludes the proof. $\square$

Proof of Theorem 2.1: If $\alpha \geq \mathsf{P}(|\hat{\vartheta}_0 - \theta| > h)$ then stopping at $T_o = 0$ corresponds to the smallest possible (average) number of samples while, at the same time, we satisfy the coverage probability constraint.

To prove ii) we first show that there exists N_α such that $\mathsf{P}(|\hat{\vartheta}_{N_\alpha} - \theta| > h) < \alpha$. Note that

$$\mathsf{P}(|\hat{\vartheta}_t - \theta| > h) \leq \mathsf{P}\left(\left|\frac{S_t}{t} - \theta\right| > h\right) \leq \frac{1}{h^2}\mathsf{E}\left[\left(\frac{S_t}{t} - \theta\right)^2\right]$$
$$= \frac{1}{h^2}\mathsf{E}\left[\mathsf{E}\left[\left(\frac{S_t}{t} - \theta\right)^2 \middle|\theta\right]\right] = \frac{1}{h^2}\mathsf{E}\left[\frac{\theta(1-\theta)}{t}\right] \leq \frac{1}{4h^2t}, \quad (2.43)$$

where we used the fact that $\frac{S_t}{t}$ is not the optimum Bayes estimator of the mid-point, then we applied the Chebyshev's inequality, then the fact that $\frac{S_t}{t}$ is an estimator of θ with estimation error variance equal to $\frac{\theta(1-\theta)}{t}$ and finally that $\theta(1-\theta) \leq \frac{1}{4}$. From (2.43) we conclude that $\mathsf{P}(|\hat{\vartheta}_t - \theta| > h) \to 0$ as $t \to \infty$ therefore, there exists N_α such that $\mathsf{P}(|\hat{\vartheta}_{N_\alpha} - \theta| > h) < \alpha$.

Fix $N \geq N_\alpha$ and denote $\mathscr{V}_t(S_t, c) = \inf_{t \leq T \leq N} \mathsf{E}[c(T-t) + \mathscr{C}_T|\mathscr{F}_t]$, where we underline the dependence of $\mathscr{V}_t$ on c (in addition to S_t). For $0 \leq c_1 \leq c_2$ and $T \geq t$ we can write

$$c_1(T-t) + \mathscr{C}_T \leq c_2(T-t) + \mathscr{C}_T,$$

which, after taking expectation conditioned on $\mathscr{F}_t$ and then infimum over $t \leq T \leq N$, proves that $\mathscr{V}_t(S_t, c)$ is increasing in c. The increase of $\mathscr{V}_t(S_t, c)$ with respect to c also suggests that the optimum stopping time $T_{\rm o}(c)$, defined in (2.21), is a decreasing function of c.

Consider now the sequence of optimum complementary coverage probabilities $\{\mathscr{C}_t\}$, we observe

$$\mathscr{C}_t = \inf_{\hat{\theta}} \mathsf{P}(|\hat{\theta} - \theta| > h|\mathscr{F}_t) = \inf_{\hat{\theta}} \mathsf{E}[\mathsf{P}(|\hat{\theta} - \theta| > h|\mathscr{F}_{t+1})|\mathscr{F}_t]$$
$$\geq \mathsf{E}\left[\inf_{\hat{\theta}} \mathsf{P}(|\hat{\theta} - \theta| > h|\mathscr{F}_{t+1})|\mathscr{F}_t\right] = \mathsf{E}[\mathscr{C}_{t+1}|\mathscr{F}_t]. \quad (2.44)$$

We can then write

$$\mathsf{P}(|\hat{\vartheta}_{T_{\rm o}(c)}-\theta|>h)=\mathsf{E}[\mathscr{C}_{T_{\rm o}(c)}]=\mathscr{C}_0-\mathsf{E}\left[\sum_{t=0}^{T_{\rm o}(c)-1}\{\mathscr{C}_t-\mathscr{C}_{t+1}\}\right]$$

$$=\mathscr{C}_0-\mathsf{E}\left[\sum_{t=0}^{N}\{\mathscr{C}_t-\mathscr{C}_{t+1}\}\mathbb{1}\{T_{\rm o}(c)>t\}\right]=\mathscr{C}_0-\mathsf{E}\left[\sum_{t=0}^{N}\{\mathscr{C}_t-\mathsf{E}[\mathscr{C}_{t+1}|\mathscr{F}_t]\}\mathbb{1}\{T_{\rm o}(c)>t\}\right],$$

where for the last equality we used the fact that $\mathbb{1}\{T_{\rm o}(c)>t\}$ is $\mathscr{F}_t$-measurable. This combined with (2.44) and the decrease of $T_{\rm o}(c)$ with respect to c, implies that $\mathsf{P}(|\hat{\vartheta}_{T_{\rm o}(c)}-\theta|>h)$ is increasing in c.

For $c=1$ we stop at 0 and, therefore, $\mathsf{P}(|\hat{\vartheta}_{T_{\rm o}(1)}-\theta|>h)=\mathsf{P}(|\hat{\vartheta}_0-\theta|>h)>\alpha$. Set now $c=0$ which suggests that the cost of sampling is zero and therefore the optimum is to stop at N (we also deduce this by combining (2.21) and (2.44)). This yields $\mathsf{P}(|\hat{\vartheta}_{T_{\rm o}(0)}-\theta|>h)=\mathsf{P}(|\hat{\vartheta}_N-\theta|>h)=\mathsf{E}[\mathscr{C}_N]$. Now from (2.44) by averaging we conclude that $\mathsf{E}[\mathscr{C}_t]$ is decreasing in t and for $N>N_\alpha$ we have $\mathsf{E}[\mathscr{C}_N]\leq\mathsf{E}[\mathscr{C}_{N_\alpha}]<\alpha$, implying $\mathsf{P}(|\hat{\vartheta}_{T_{\rm o}(0)}-\theta|>h)<\alpha$. As mentioned, $\mathsf{P}(|\hat{\vartheta}_{T_{\rm o}(c)}-\theta|>h)$ is increasing in c, if it is also continuous then there exists $0<c_*<1$ satisfying $\mathsf{P}(|\hat{\vartheta}_{T_{\rm o}(c_*)}-\theta|>h)=\alpha$ which means that $T_{\rm o}(c_*)$ solves the constrained problem. In case the function $\mathsf{P}(|\hat{\vartheta}_{T_{\rm o}(c)}-\theta|>h)$ exhibits a jump at c_* such that for c_*- the probability is strictly smaller than α while for c_*+ it is strictly larger, then before taking any samples we need to perform a randomization to decide which of the two stopping times $T_{\rm o}(c_*\text{-}),T_{\rm o}(c_*\text{+})$ to use. The randomization probability must be selected so that we satisfy the constraint with equality. $\square$

Proof of Lemma 2.4: We prove (2.27) first. Set $\Omega_N=\varnothing$, i.e. we stop necessarily at N. Then we note that

$$T=\sum_{t=0}^{N-1}\mathbb{1}\{T>t\}$$

$$=(1+\cdots(1+\mathbb{1}\{S_{N-1}\in\Omega_{N-2}\}(1+\mathbb{1}\{S_{N-1}\in\Omega_{N-1}\}(1+\mathbb{1}\{S_N\in\Omega_N\})))\cdots)$$

suggesting that

$$\mathsf{E}[T|\theta] = \mathsf{E}[(1+\cdots \mathsf{E}[(1+\mathbb{1}\{S_{N-1} \in \Omega_{N-1}\}\mathsf{E}[(1+\mathbb{1}\{S_N \in \Omega_N\})|\mathscr{F}_{N-1}, \theta)|\mathscr{F}_{N-2}, \theta]\cdots)|\theta].$$

If we set $U_N(S_N) = 0$ then we can define the backward recursion

$$\begin{aligned} U_t(S_t) = \mathsf{E}[1+\mathbb{1}\{S_{t+1} \in \Omega_{t+1}\}U_{t+1}(S_{t+1})|\mathscr{F}_t] &= 1+\mathsf{E}[\mathbb{1}\{S_{t+1} \in \Omega_{t+1}\}U_{t+1}(S_{t+1})|\mathscr{F}_t] \\ &= 1 + \mathsf{P}(X_{t+1} = 1|S_t, \theta)\mathbb{1}\{S_t + 1 \in \Omega_{t+1}\}U_{t+1}(S_t + 1) \\ &\quad + \mathsf{P}(X_{t+1} = 0|S_t, \theta)\mathbb{1}\{S_t \in \Omega_{t+1}\}U_{t+1}(S_t) \\ &= 1 + \theta\mathbb{1}\{S_t + 1 \in \Omega_{t+1}\}U_{t+1}(S_t + 1) + (1-\theta)\mathbb{1}\{S_t \in \Omega_{t+1}\}U_{t+1}(S_t), \end{aligned}$$

which proves (2.27) and, also, that $U_0(S_0) = \mathsf{E}[T|\theta]$. For (2.28) we proceed similarly the only difference being that $\mathsf{P}(X_{t+1} = 1|S_t) = g_{t+1}(S_t)$ with this probability being defined in (2.20).

For (2.29) and (2.30) we follow similar steps. We have

$$\begin{aligned} \mathbb{1}\{|\hat{\theta}_T - \theta| > h\} &= \sum_{t=0}^{N} \mathbb{1}\{|\hat{\theta}_t - \theta| > h\}\mathbb{1}\{T = t\} \\ &= \sum_{t=0}^{N} \mathbb{1}\{|\hat{\theta}_t - \theta| > h\}\mathbb{1}\{S_t \notin \Omega_t\}\prod_{j=0}^{t-1}\mathbb{1}\{S_j \in \Omega_j\} \\ &= (\mathbb{1}\{|\hat{\theta}_0 - \theta| > h\}\mathbb{1}\{S_0 \notin \Omega_0\}) + (\mathbb{1}\{|\hat{\theta}_1 - \theta| > h\}\mathbb{1}\{S_1 \notin \Omega_1\})\mathbb{1}\{S_0 \in \Omega_0\} + \cdots \\ &\quad + (\mathbb{1}\{|\hat{\theta}_N - \theta| > h\}\mathbb{1}\{S_N \notin \Omega_N\})\prod_{j=1}^{N-1}\mathbb{1}\{S_j \in \Omega_j\}. \end{aligned}$$

Applying expectation given θ yields

$$\begin{aligned}
&\mathsf{P}(|\hat{\theta}_T - \theta| > h|\theta) \\
&\qquad = \mathsf{E}[\mathbb{1}\{|\hat{\theta}_0 - \theta|\}\mathbb{1}\{S_0 \notin \Omega_0\} + \cdots + \mathsf{E}[\mathbb{1}\{|\hat{\theta}_{N-1} - \theta| > h\}\mathbb{1}\{S_{N-1} \notin \Omega_N\} \\
&\quad + (\mathsf{E}[\mathbb{1}\{|\hat{\theta}_N - \theta| > h\}\mathbb{1}\{S_N \notin \Omega_N\}|\mathscr{F}_{N-1}, \theta])\mathbb{1}\{S_{N-1} \in \Omega_{N-1}\}|\mathscr{F}_{N-2}, \theta]) \cdots |\theta].
\end{aligned}$$

Defining $W_N(S_N) = \mathbb{1}\{|\hat{\theta}_N - \theta| > h\}$ it is straightforward to see that the recursion in (2.29) computes the desired complementary coverage probability. Similarly for (2.30) only now instead of conditioning with respect to both $\mathscr{F}_t$ and θ we condition only with respect to $\mathscr{F}_t$. This concludes the proof. $\square$

Proof of Theorem 2.2: Let us first find upper and lower bounds of $\mathscr{C}_t(S_t)$ that are independent from S_t. From [46, Theorem 2.1] and for a random variable X with density Beta(x, p, q) we have that

$$\mathsf{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2}{8(p+q+1)}},\ \lambda > 0, \tag{2.45}$$

where $\mu = \frac{p}{p+q}$ is the average under the Beta density. Using the Markov inequality we can then write

$$\begin{aligned}
\mathsf{P}(|X - \mu| > h) &= \mathsf{P}(X - \mu > h) + \mathsf{P}(X - \mu < -h) \\
&= \mathsf{P}(X - \mu > h) + \mathsf{P}(1 - X - (1 - \mu) > h) \\
&\leq \frac{\mathsf{E}[e^{\lambda(X-\mu)}]}{e^{\lambda h}} + \frac{\mathsf{E}[e^{\lambda(1-X-(1-\mu))}]}{e^{\lambda h}} \leq 2e^{\frac{\lambda^2}{8(p+q+1)} - \lambda h},
\end{aligned} \tag{2.46}$$

where we used the fact that if X is Beta distributed with parameters p, q then $1 - X$ is also Beta with parameters q, p. Selecting in (2.46) $\lambda = 4(p + q + 1)h$ yields the tightest upper bound, namely

$$\mathsf{P}(|X - \mu| > h) \leq 2e^{-2h^2(p+q+1)}. \tag{2.47}$$

We can now use this result to upper bound $\mathscr{C}_t(S_t)$. We observe that

$$\mathscr{C}_t(S_t) = \inf_{\hat{\theta}_t} \mathsf{P}(|\theta - \hat{\theta}_t| > h|\mathscr{F}_t) \leq \mathsf{P}\big(|\theta - \mathsf{E}[\theta|\mathscr{F}_t]| > h|\mathscr{F}_t\big) \leq 2e^{-2h^2(t+2a+1)}. \quad (2.48)$$

For for the last inequality we used (2.47) and the fact that θ given $\mathscr{F}_t$ is Beta distributed with parameters $p = S_t + a$ and $q = t - S_t + a$.

Let us now find a lower bound for $\mathscr{C}_t(S_t)$. From [28, Page 944, Formula 26.5.15] we conclude that $I_x(p,q) > I_x(p+1, q-1)$ for $q > 1$. Using this inequality repeatedly in (2.32) we conclude

$$\begin{aligned}\mathsf{P}(|\hat{\theta}_t - \theta| > h|\mathscr{F}_t) &= 1 - I_{\min\{1,\hat{\theta}_t+h\}}(S_t + a, t - S_t + a) + I_{\max(0,\hat{\theta}_t-h)}(S_t + a, t - S_t + a)\\ &= I_{\max\{0,1-h-\hat{\theta}_t\}}(t - S_t + a, S_t + a) + I_{\max(0,\hat{\theta}_t-h)}(S_t + a, t - S_t + a)\\ &\geq I_{\max\{0,1-h-\hat{\theta}_t\}}(t + 2n_a + \delta_a, \delta_a) + I_{\max\{0,\hat{\theta}_t-h\}}(t + 2n_a + \delta_a, \delta_a), \quad (2.49)\end{aligned}$$

where for the second equality we used the property $1 - I_x(p,q) = I_{1-x}(q,p)$ and where n_a, δ_a are defined as

$$n_a = \begin{cases} [a] & \text{if } a \text{ not an integer} \\ a-1 & \text{if } a \text{ an integer,} \end{cases} \qquad \delta_a = \begin{cases} a-[a] & \text{if } a \text{ not an integer} \\ 1 & \text{if } a \text{ an integer,} \end{cases}$$

where $[a]$ denotes integer part of a. Since $a > 0$ we have $n_a \geq 0$, $1 \geq \delta_a > 0$ and $a = n_a + \delta_a$. By taking the derivative of the last sum in (2.49) with respect to $\hat{\theta}_t$ we can show that it has the same sign as the following expression

$$\phi(\hat{\theta}_t) = \frac{(\hat{\theta}_t - h)^{t+2n_a+\delta_a-1}}{(1 + h - \hat{\theta}_t)^{1-\delta_a}} - \frac{(1 - h - \hat{\theta}_t)^{t+2n_a+\delta_a-1}}{(\hat{\theta}_t + h)^{1-\delta_a}}.$$

Now it is easy to verify that $\phi(1 - \hat{\theta}_t) = -\phi(\hat{\theta}_t)$ therefore it is sufficient to analyze the sign of $\phi(\hat{\theta}_t)$ for $h \leq \hat{\theta}_t \leq 0.5$. When $t \geq 1$ and because $1 \geq \delta_a$ we can see that the sign is negative for any value of a, suggesting that we have a minimum for $\hat{\theta}_t = 0.5$. Therefore, if $\Gamma(x)$ denotes the Gamma function, then for $t \geq 1$ we can write

$$\begin{aligned}\mathscr{C}_t \geq 2I_{0.5-h}(t+2n_a+\delta_a,\delta_a) \geq 2\frac{\Gamma(t+2n_a+2\delta_a)(0.25-h^2)^{\delta_a}}{\Gamma(t+2n_a+\delta_a+1)\Gamma(\delta_a)}(0.5-h)^{t+2n_a}\\ = 2\frac{\Gamma(t+2n_a+2\delta_a+1)(0.25-h^2)^{\delta_a}}{(t+2n_a+2\delta_a)\Gamma(t+2n_a+\delta_a+1)\Gamma(\delta_a)}(0.5-h)^{t+2n_a}\\ \geq 2\frac{(0.25-h^2)^{\delta_a}}{(t+2n_a+2\delta_a)\Gamma(\delta_a)}(0.5-h)^{t+2n_a}. \quad (2.50)\end{aligned}$$

In the previous expression the second inequality comes from [28, Page 944, Formula 26.5.16]; for the next equality we used the property $\Gamma(x+1) = x\Gamma(x)$; while for the last inequality we used the increase of $\Gamma(x)$ for $x \geq 1.5$, which is true in our case for $t \geq 1$ and any $a > 0$.

Having established bounds for $\mathscr{C}_t$ we can now compute an upper bound N for t_{up} and a lower bound ν for t_{lo} therefore proving their existence and demonstrating properties i) and ii). We first note that if $\mathscr{C}_N \leq c$ in (2.18) we will have $\mathscr{C}_N \leq c + \tilde{\mathscr{V}}_N$ meaning that $\mathscr{V}_N = \mathscr{C}_N$ and consequently N is a stopping instant for *all* values of S_t. This implies that $T_{\text{o}} \leq N$. Quantity t_{up} is the smallest N for which this inequality is true for all S_t. Requiring $2e^{-2h^2(N+2a+1)} \leq c$ we obtain

$$N = \left\lceil \max\left\{0, \frac{|\log(c)| + \log(2)}{2h^2} - 2a - 1\right\}\right\rceil. \quad (2.51)$$

To find a lower bound ν for t_{lo} we combine the lower bound of $\mathscr{C}_t$ with an upper bound for $\mathscr{V}_t$. Finding the latter is straightforward. Indeed if we start from time instant N which, as we argued, is selected so that $\mathscr{C}_N \leq c$, then using induction and the fact that

$$\mathscr{V}_t = \min\{\mathscr{C}_t, c + \mathsf{E}[V_{t+1}|\mathscr{F}_t]\} \leq c + \mathsf{E}[\mathscr{V}_{t+1}|\mathscr{F}_t]$$

we can show that $\mathscr{V}_t \leq c + c(N-t) = c(N+1-t)$. It is then clear that, as long as

$c(N+1) \leq \mathscr{C}_0$, for any $t \geq 1$ for which we have

$$c(N+1-t)(t+2n_a+2\delta_a) \leq 2\frac{(0.25-h^2)^{\delta_a}}{\Gamma(\delta_a)}(0.5-h)^{t+2n_a} \tag{2.52}$$

we do not stop at this time instant. In fact we can see that we have an interval of the form $t \in [0, \ldots, \nu]$ during which no stopping can occur. A rough estimate of ν can be obtained by solving instead of (2.52) the simpler alternative $\max_t c(N+1)(t+2n_a+2\delta_a) = \frac{c}{4}(N+2n_a+2\delta_a)^2 \leq \frac{2(0.25-h^2)^{\delta_a}}{\Gamma(\delta_a)}(0.5-h)^{\nu+2n_a}$ which yields

$$\nu = \left\lfloor \max\left\{0, \frac{|\log(c)| - \log\left((N+2n_a+\delta_a)^2\Gamma(\delta_a)\right) + \log\left(8(0.25-h^2)^{\delta_a}\right)}{|\log(0.5-h)|} - 2n_a\right\}\right\rfloor, \tag{2.53}$$

provided c satisfies $c \leq \frac{\mathscr{C}_0}{N+1}$. Regarding the latter, if we are in the non-trivial case where we do not stop at time 0 then $\alpha < \mathscr{C}_0$, consequently it is sufficient to have $c \leq \frac{\alpha}{N+1}$. We thus conclude that for small enough c there is a lower limit $t_{\mathrm{lo}} \geq \nu$ which is nontrivial. This concludes the proof. □

Acknowledgments

This work was supported by the US National Science Foundation under Grant CIF 1513373 through Rutgers University and under Grant CMMI 1362876 through Georgia Institute of Technology.

CHAPTER 3

TANDEM-WIDTH SEQUENTIAL CONFIDENCE INTERVALS FOR A BERNOULLI PROPORTION

3.1 Introduction

In this chapter, we investigate a sequential CI of a Bernoulli proportion θ but with a new twist of *tandem-width*. By *tandem-width*, we mean that the half-width h of the $100(1-\alpha)\%$ CI is not fixed beforehand; it is instead required to satisfy two different upper bounds, h_0 and h_1, depending on the values of θ. Our motivating examples are the customer click-through rate to measure the efficacy of a new online ad marketing campaign, and the Statistical Model Checking (SMC) approach in complicated stochastic systems, e.g., see Jegourel, Sun, and Dong [20]. In both of these modern applications, it is very expensive and time-consuming to set up the experiments or simulations. Once they are set up, one wants to use the least amount of samples to gain knowledge of the Bernoulli proportion θ as *accurately and precisely* as possible due to the time or cost constraints. For instance, for the 95% CI of θ, if the true θ is estimated to be in the interval $[0.2, 0.8]$, one may feel that the half-width $h_0 = 0.1$ is precise enough and is acceptable. On the other hand, if the true θ is in the interval $(0, 0.1)$ or $(0.9, 1)$, one may feel that $h_0 = 0.1$ is too crude, and the half-width $h_1 = 0.01$ might be more suitable. This inspires us to investigate the problem of tandem-width sequential interval estimation.

We propose to develop effective sequential methods for tandem-width interval estimation of the Bernoulli proportion θ at the pre-specified confidence level $100(1-\alpha)\%$. It is intuitive to combine two sequential fixed-width CI together, one for each fixed-width, but the difficulty is that these sequential fixed-width CIs should be not only statistically efficient, but also computationally efficient for the purpose of combination in practice. While many existing methods can be statistically efficient to derive a $100(1-\alpha)\%$ CI in the sense of small expected sample sizes for a fixed-width h, the stopping time $T(h)$ will often heavily depend on the fixed-width $h.$ In particular, it is unclear whether the stopping boundary of $T(h)$ at each time step is a monotone function of h or not, and thus it is not easy to implement the combination of two different stopping times $T(h)$'s. As a concrete illustration, the stopping time of the sequential CI proposed by Frey [23] is based on the Bayesian point estimator, which depends on the half-width h when the prior distribution is optimized for the smallest expected sample size. To circumvent this problem, we propose to use the minimax point estimator of the Bernoulli proportion to develop effective sequential methods for fixed-width sequential CI, where the decision statistics do not depend on the half-width h, which only affects the stopping boundaries of the detection statistics monotonically. This allows us to conveniently combine two fixed-width sequential interval estimation methods together to derive an efficient tandem-width sequential interval estimation method.

The remainder of this article is organized as follows. In Section 3.2, we formulate our problem on tandem-width sequential CIs for a Bernoulli proportion θ and provide some background regarding different point estimators for θ and, in particular, on the method proposed by Frey [23]. In Section 3.3, we discuss our sequential stopping rules for the fixed-width CI and the tandem-width CI, and show some asymptotic properties for our proposed methods. Section 3.4 presents simulation results for our tandem-width stopping rule. Moreover, we also provide numerical results that compare our proposed fixed-width stopping rule to Frey's stopping rule. The concluding remarks are included in Section 3.5.

3.2 Problem Formulation and Background

Assume that we observe a sequence of i.i.d. Bernoulli random variables, $X_1, X_2, \cdots$ sequentially, i.e., one at a time. Suppose $\mathsf{P}(X_i = 1) = \theta$ and $\mathsf{P}(X_i = 0) = 1 - \theta$, and we want to use as few samples as possible to make an accurate interval estimation about the unknown parameter $\theta \in [0, 1]$ at the confidence level $100(1 - \alpha)\%$ for some pre-specified

α, say, $\alpha = 5\%$. Since an interval $[a, b]$ can be rewritten as the form of $[c - h, c + h]$ with $c = (a+b)/2$ and $h = (b-a)/2$, below we assume that the $100(1-\alpha)\%$ CI of θ is written in the form of $[\hat{\theta}_T - h, \hat{\theta}_T + h]$, where h is the desired half-width of the CI, and $\hat{\theta}_T$ can be thought as the point estimator of θ when we stop taking observations at time T.

In the problem of tandem-width sequential confidence intervals, we want to find a stopping time T and when stopped, we are able to derive a $100(1-\alpha)\%$ CI of θ whose half-width is required to satisfy two different upper bounds, h_0 and h_1, depending on the point estimate of θ. To be more concrete, assume that the derived sequential $100(1-\alpha)\%$ CI for θ is of the form $[\hat{\theta}_T - h, \hat{\theta}_T + h]$, where $\hat{\theta}_T$ is the point estimator of θ when stopping taking observations at time T. On the one hand, when the estimate $\hat{\theta}_T$ is not too small or large, say, when $\hat{\theta}_T \in [\theta_0, 1-\theta_0]$ for some pre-specified θ_0, e.g., $\theta_0 = 0.1$, we would like to set the half-width h of the CI to be a relatively large value h_0 (e.g., $h_0 = 0.1$) so as to save time and sampling costs. On the other hand, when $\hat{\theta}_T$ is quite small or large, say, when $\hat{p}_T \le \theta_0$ or $\ge 1-\theta_0$, we would like to set the half-width of the CI to be a smaller value h_1 (e.g., $h_1 = 0.01$) in order for the CI to be more meaningful. In other words, in the latter case, it is more useful to take a longer time to derive a meaningful CI instead of stopping earlier to derive a meaningless CI.

To be more rigorous, we would like to find a stopping time T and the corresponding estimator $\hat{\theta}_T$ that minimize the average run lengths (ARLs), $\mathsf{E}_\theta(T)$, simultaneously for all $0 \le \theta \le 1$, subject to the coverage probability (CP) constraint that

$$\begin{aligned}
\mathrm{CP}_\theta(h_0) &= \mathsf{P}_\theta\left(\theta \in [\hat{\theta}_T - h_0, \hat{\theta}_T + h_0]\right) \ge 1-\alpha, \quad \text{when } \theta_0 \le \theta \le 1-\theta_0, \text{ and} \\
\mathrm{CP}_\theta(h_1) &= \mathsf{P}_\theta\left(\theta \in [\hat{\theta}_T - h_1, \hat{\theta}_T + h_1]\right) \ge 1-\alpha, \quad \text{when } \theta < \theta_0 \text{ or } \theta > 1-\theta_0. \quad (3.1)
\end{aligned}$$

where $0 < h_1 < h_0 < 1$ and $\alpha > 0$ are pre-specified, e.g., $h_0 = 0.1, h_1 = 0.01$ and $\alpha = 5\%$.

Let us now provide some background information on the point and interval estimation of Bernoulli proportion θ. For this purpose, we first review three different kinds of point estimators of θ under the offline context when the complete observations are $\{X_1, X_2, \cdots, X_t\}$: maximum likelihood estimator (MLE), Bayes estimator, and minimax estimator, denoted by $\hat{\theta}_t, \tilde{\theta}_t, \theta_t^{\star}$ below, respectively, to emphasize their dependence on the sample size t. To this end, assume that $X_1, X_2, \ldots, X_t$ are i.i.d. Bernoulli(θ). The MLE of θ is the sample mean

$$\hat{\theta}_t = S_t/t \qquad \text{where} \qquad S_t = \sum_{i=1}^{t} X_i. \tag{3.2}$$

Below we follow the literature to assume that the point estimator $\hat{\theta}_T$ in (3.1) is the MLE estimator in (3.2) when stopping taking observations at T. This will allow us to have a fair comparison between our proposed stopping time T with other sequential methods in the literature.

As for the Bayes estimator of θ, it is well-known that if the prior distribution of θ is the Beta(α, β) distribution for some pre-specified $\alpha, \beta > 0$, then the posterior of θ given observed $(X_1, X_2, \ldots, X_t)$ is a Beta($\alpha + S_t, \beta - S_t + t$) distribution. Thus the mean of the posterior distribution, $(S_t + \alpha)/(t + \alpha + \beta)$, is the Bayes estimator of θ under the standard square error loss function. One important special case of the prior Beta distribution is when $\alpha = \beta = a$ for some $a > 0$, so that the corresponding Bayes estimator of θ becomes

$$\tilde{\theta}_{t,a} = (S_t + a)/(t + 2a). \tag{3.3}$$

Meanwhile, under the squared error loss function, the minimax framework is to find an estimator $\delta = \delta(X_1, \cdots, X_t)$ that minimizes the largest mean square error over the whole space $[0, 1]$ of the true parameter θ. In other words, the minimax estimator minimizes $\max_{0 \leq \theta \leq 1} \mathsf{E}_{\theta}[(\delta - \theta)^2]$, where E_{θ} denotes the expectation when θ is the true Bernoulli parameter. For Bernoulli random variables and for fixed sample size t, the minimax estimator is known to be given by

$$\theta_t^\star = (S_t + \sqrt{t}/2)/(t + \sqrt{t}), \tag{3.4}$$

see, for example, Lehmann and Casella [32, pp. 311–312]. Note that $\theta_t^\star$ is minimax and not asymptotically minimax because it is Bayes with respect to the least favorable prior distribution Beta($\sqrt{t}/2, \sqrt{t}/2$) and has a constant risk or mean square error of $t/(4(t + \sqrt{t})^2)$.

It is useful to compare the Bayes estimator $\tilde{\theta}_{t,a}$ in (3.3) with the minimax estimator $\theta_t^\star$ in (3.4). On the one hand, for a fixed sample size n, the minimax estimator $\theta_t^\star$ can be thought of as a special case of the Bayes estimator with $a = \sqrt{t}/2$. On the other hand, when the sample size t is variable, the estimators are fundamentally different: the minimax estimator incorporates the sample size t adaptively in the estimator itself, whereas the Bayes estimator involves a constant parameter a that can be tuned for optimization depending on the problem context.

Next, we review the well-known off-line sample size formula for estimating Bernoulli proportion θ. Recall that in the offline context with a fixed sample size t, by the Central Limit Theorem (CLT), we have $(\hat{\theta}_t - \theta)/\sqrt{\theta(1-\theta)/t} \sim N(0,1)$ for large t, where $\hat{\theta}_t$ is the MLE in (3.2). Thus an (approximate) $100(1-\alpha)\%$ CI of θ is $\hat{\theta}_t \pm z_{\alpha/2}\sqrt{\theta(1-\theta)/t}$. If we would like the half-width of this CI to be at most h, then the sample size t needs to satisfy

$$z_{\alpha/2}\sqrt{\theta(1-\theta)/t} \le h. \tag{3.5}$$

Equivalently, the fixed-sample lower bound on the required sample size for $100(1-\alpha)$ CI is

$$t_0 = \theta(1-\theta)(z_{\alpha/2}/h)^2. \tag{3.6}$$

When we do not have any prior knowledge of θ, it is often conservative to set the conservative fixed sample size $t_{\text{cons}} = 0.25(z_{\alpha/2}/h)^2$ by using the fact that $\theta(1-\theta) \leq 0.25$ for any $0 \leq \theta \leq 1$. For instance, for the survey polls, it is typical to set $\alpha = 5\%$ and $h = 3\%$ (often called the margin of errors), and thus the conservative required sample size will be $t_{\text{cons}} = 0.25(1.96/0.03)^2 \approx 1068$.

Note that the conservative required fixed sample size t_{cons} depends heavily on the half-width h : it is $t_{\text{cons}} = 0.25(1.96/0.1)^2 \approx 97$ if half-width $h = 0.1$, but it becomes $0.25(1.96/0.01)^2 \approx 9604$ if half-width $h = 0.01$. In modern applications, smaller half-width often makes sense only when the true θ value is very small or very large, and this allows us to significantly reduce the sample size from the conservative required fixed sample size by using (3.6) when we have a prior knowledge on the bounds of θ. For instance, for the half-width $h = 0.01$, if we have prior knowledge that θ is very small or very large, say, $\theta \leq 0.03$ or $\theta \geq 0.97$, then we can significantly reduce the required sample size from the conservative value $t_{\text{cons}} = 9604$ to $t_0 = 0.03 \times 0.97 \times (1.96/0.01)^2 \approx 1118$, which is more manageable in many modern applications. This is exactly the main idea in the sequential context, where we are able to update our estimate of θ over time as we collect the data, which in turn allows us to reduce the required sample size.

Finally, let us review the existing methods for sequential fixed-width CIs for θ. The fixed-width sequential CI problem can be thought of as a special case of our proposed tandem-width CI problem when $h_0 = h_1 = h$ in (3.1). In other words, in the fixed-width $100(1-\alpha)\%$ sequential CI, we would like to find a stopping time T that minimizes the ARLs, $\mathsf{E}_\theta(T)$, simultaneously for all $0 \leq \theta \leq 1$, subject to the CP constraint that

$$\inf_{0<\theta<1} \mathsf{P}_\theta\left(\theta \in [\hat{\theta}_T - h, \hat{\theta}_T + h]\right) \geq 1 - \alpha, \tag{3.7}$$

where $\alpha > 0$ and $h > 0$ are pre-specified (e.g., $\alpha = 5\%$ and $h = 0.1$).

In the context of sequential CIs with fixed half-width h, if one stops taking observations at the stopping time T, one often writes the sequential CI as the form of $[\hat{\theta}_T - h, \hat{\theta}_T + h]$, where h is the fixed half-width and $\hat{\theta}_T$ is often the MLE in (3.2). Of course, when the lower bound $\hat{\theta}_T - h \leq 0$ or the upper bound $\hat{\theta}_T + h \geq 1$, we can threshold these values to 0 and 1, respectively, as $0 \leq \theta \leq 1$. Note that no statistical procedure can exactly and simultaneously optimize over all $0 < \theta < 1$, and thus it is reasonable to adopt the asymptotic approach as $h, \alpha \to 0$, e.g., finding a family of stopping times $T = T_{h,\alpha}$ such that $\mathsf{E}_\theta(T)$ is asymptotically equivalent to the fixed-sample lower bound in (3.6) at each $0 < \theta < 1$.

Most, if not all, existing methods for fixed-width sequential CIs are to explore the relationship (3.5) by estimating the unknown true θ carefully, especially at the early stage when few samples are available. To highlight the challenge of sequential CIs, let us estimate the unknown θ in (3.5) by the MLE $\hat{\theta}_t$ (3.2). This will yield a naive stopping time defined by

$$T_{\mathrm{W}} = \inf\{t \geq 1 : \hat{\theta}_t(1 - \hat{\theta}_t)/t \leq (h/z_{\alpha/2})^2\}. \tag{3.8}$$

Unfortunately, T_{W} in (3.8) is not efficient. In fact, when $t = 1$, the MLE $\hat{\theta}_t = 0$ or 1, thus T_{W} will always stop at time 1! There are many ways to improve this stopping time, say, implementing it only after taking $m_0 \geq 2$ observations or setting lower bounds of $\hat{\theta}_t(1 - \hat{\theta}_t)$, but the corresponding new stopping times often involve new tuning parameters and become very involved.

Frey [23] proposes an interesting idea to salvage (3.8) by using the Bayes estimate $\tilde{\theta}_{t,a}$ in (3.3), and this yields the stopping time

$$T_{\mathrm{F}} = \inf\{t \geq 1 : \tilde{\theta}_{t,a}(1 - \tilde{\theta}_{t,a})/t \leq (h/z_{\gamma/2})^2\}, \tag{3.9}$$

where the parameter $\gamma = \gamma(a, h, \alpha)$ is chosen so as to satisfy the CP constraint in (3.7). The main advantage of Frey's method T_{F} in (3.9) is that it is intuitively appealing and avoids

Table 3.1: Optimal choices of a and γ for $100(1-\alpha)\%$ CIs with fixed half-width h in Frey [23].

$1-\alpha$	90%		95%		99%	
h	a	γ	a	γ	a	γ
0.10	4	0.0754	4	0.0356	6	0.0068
0.05	4	0.0859	6	0.0433	8	0.0083
0.01	8	0.0972	10	0.0487	14	0.0097

the trivial stopping scenario of (3.8). Unfortunately, in Frey's method T_{F} in (3.9), both the decision statistics and the threshold $(h/z_{\gamma/2})^2$ depends on the tuning parameter a, which needs to be optimized according to the specific half-width h and confidence level $1-\alpha$, see Table 3.1 for the optimal values of a and γ recommended by Frey [23]. As a result, it is challenging to combine two fixed-width sequential CIs derived by Frey's method together in the tandem-width sequential CI context.

3.3 Proposed Sequential Methods

In the problem of tandem-width sequential CI, we propose to develop sequential methods by combining two efficient sequential methods that are designed for fixed-width CIs. For efficiency and easy implementation, we require that these two sequential methods for fixed-width CIs have the same decision statistics, with the only difference being the thresholds of the decision statistics. For this purpose, we propose to use the minimax estimator $p_t^{\star}$ in (3.4) to estimate the unknown θ in the variance estimation in (3.5). This allows us to develop effective stopping times that do not involve tuning parameters.

To better present our proposed methods, this section is divided into three parts: Subsection 3.3.1 presents our proposed stopping times for sequential CIs; Subsection 3.3.2 discusses the asymptotic properties of our proposed sequential methods; and finally, Subsection 3.3.3 discusses the finite-sample numerical issues. Specifically, in Subsection 3.3.3 we discuss how to accurately compute the ARL and CP properties of our proposed sequential methods by non-Monte-Carlo numerical methods. This will allow us to validate our theoretical results.

3.3.1 Proposed stopping times

To simplify our notation, below we fix the α value in the CP constraint in (3.7), and write our proposed stopping times as a function of half-width h of the CI.

Let us begin with our proposed stopping time for sequential $100(1-\alpha)\%$ CI with the fixed half-width h. Our key idea is to apply the minimax estimator $\theta_t^\star$ in (3.4) to estimate θ in (3.5). This motivates us to propose the following stopping time:

$$T_{\mathrm{M}}(c) = \inf\{t \geq 1 : \theta_t^\star(1-\theta_t^\star)/t \leq c\}, \tag{3.10}$$

where the threshold $c = c_h$ is chosen to satisfy the CP constraint in (3.7). When our proposed stopping time $T_{\mathrm{M}}(c)$ in (3.10) stops, we will report the fixed-width sequential CI of θ as $[\hat{\theta}_{T_{\mathrm{M}}(c)} - h, \hat{\theta}_{T_{\mathrm{M}}(c)} + h]$, or more accurately as $[\max(0, \hat{\theta}_{T_{\mathrm{M}}(c)} - h), \min(1, \hat{\theta}_{T_{\mathrm{M}}(c)} + h)]$.

It is important to point out that the threshold $c = c_h$ in (3.10) is an increasing function of the half-width h. To see this, note that

$$\mathsf{P}_\theta\left(\theta \in [\hat{\theta}_{T_{\mathrm{M}}(c)} - h, \hat{\theta}_{T_{\mathrm{M}}(c)} + h]\right) = \mathsf{P}_\theta\left(\hat{\theta}_{T_{\mathrm{M}}(c)} \in [\theta - h, \theta + h]\right),$$

and thus the CP constraint in (3.7) implies that $\hat{\theta}_{T_{\mathrm{M}}(c)}$ needs to be closer to the true θ with high probability for a smaller half-width h. This can only happen if the sample size $T_{\mathrm{M}}(c)$ becomes larger. Meanwhile, the stopping time $T_{\mathrm{M}}(c)$ in (3.10) or the (expected) sample size is clearly increasing as the threshold $c = c_h$ decreases. This implies that c_h is increasing in h.

For the purpose of comparison with relation (3.5) and Frey's method in (3.9), we may rewrite the threshold c in (3.10) as the form of

$$c = c_h = (h/z_{\gamma/2})^2, \tag{3.11}$$

where $\gamma = \gamma(h, \alpha)$ depends on both h and α. In the finite-sample setting, we usually have $0 < \gamma < \alpha$ due to the repeated significance tests over time in (3.10), although asymptotically $\gamma/\alpha \to 1$ as $h \to 0$, see Theorem 3.1 in the next section. Also our extensive numerical experiments suggest that γ is decreasing as a function of h, see Table 3.2 below, but unfortunately we are unable to prove it rigorously.

Now we are ready to present our proposed tandem-width sequential CI. Denote by $T_{\mathrm{M}}(c_0)$ and $T_{\mathrm{M}}(c_1)$ the stopping time $T_{\mathrm{M}}(c)$ in (3.10) with $h = h_0$ (e.g., $= 0.1$) and $h = h_1$ (e.g., $= 0.01$), respectively. Furthermore, based on the values of h_0 and h_1, we can write

$$c_0 = c_{h_0} = (h_0/z_{\gamma_0/2})^2 \text{ and } c_1 = c_{h_1} = (h_1/z_{\gamma_1/2})^2 \tag{3.12}$$

where $\gamma_0 = \gamma(h_0, \alpha)$ and $\gamma_1 = \gamma(h_1, \alpha)$. At a high-level, our proposed stopping time is a two-stage procedure: the first-stage uses our stopping time $T_{\mathrm{M}}(c_0)$ to derive a sequential CI with a larger half-width h_0, and if the estimate $\hat{\theta}_{T_{\mathrm{M}}(c_0)}$ at the end of the first-stage is too small or too large, then we continue to conduct the second-stage by using $T_{\mathrm{M}}(c_1)$ to derive another sequential CI with a smaller half-width h_1. Note that $T_{\mathrm{M}}(c_0) \leq T_{\mathrm{M}}(c_1)$, and the observations in the first-stage are kept and used in $T_{\mathrm{M}}(c_1)$ in the second-stage.

Mathematically, our proposed stopping time, denoted by T_{TW}, for the tandem-width sequential CI is defined by

$$T_{\mathrm{TW}} = \begin{cases} T_{\mathrm{M}}(c_0), & \text{if } \hat{\theta}_{T_{\mathrm{M}}(c_0)} \in [\theta_0, 1 - \theta_0]; \\ T_{\mathrm{M}}(c_1), & \text{otherwise.} \end{cases} \tag{3.13}$$

When $T_{\mathrm{TW}} = T_{\mathrm{M}}(c_0)$, we have $\hat{\theta}_{T_{\mathrm{M}}(c_0)} \in [\theta_0, 1-\theta_0]$, and thus we report the $100(1-\alpha)\%$ CI as the one with a larger half-width h_0, i.e., $[\max(0, \hat{\theta}_{T_{\mathrm{M}}(c_0)} - h_0), \min(1, \hat{\theta}_{T_{\mathrm{M}}(c_0)} + h_0)]$. When $T_{\mathrm{TW}} = T_{\mathrm{M}}(c_1)$, we have $\hat{\theta}_{T_{\mathrm{M}}(c_0)} \notin [\theta_0, 1 - \theta_0]$, and thus we generally report $100(1-\alpha)\%$ CI as the one with a smaller half-width h_1, i.e., $[\max(0, \hat{\theta}_{T_{\mathrm{M}}(c_1)} - h_1), \min(1, \hat{\theta}_{T_{\mathrm{M}}(c_1)} + h_1)]$. In the finite-sample setting, it is possible, though very rare, that $\hat{\theta}_{T_{\mathrm{M}}(c_0)} \notin [\theta_0, 1 - \theta_0]$

but $\hat{\theta}_{T_{\mathrm{M}}(c_1)} \in [\theta_0, 1-\theta_0]$. In such rare cases, when $T_{\mathrm{TW}} = T_{\mathrm{M}}(c_1)$, one may choose to report the $100(1-\alpha)\%$ CI by using $\hat{\theta}_{T_{\mathrm{M}}(c_1)}$ with a larger half-width h_0, e.g., report CI as $[\max(0, \hat{\theta}_{T_{\mathrm{M}}(c_1)} - h_0), \min(1, \hat{\theta}_{T_{\mathrm{M}}(c_1)} + h_0)]$.

For the purpose of numerical computations, it is useful to rewrite T_{TW} in (3.13) as

$$T_{\mathrm{TW}} = T_{\mathrm{M}}(c_1) - (T_{\mathrm{M}}(c_1) - T_{\mathrm{M}}(c_0)) \cdot \mathbb{1}\{\hat{\theta}_{T_{\mathrm{M}}(c_0)} \in [\theta_0, 1-\theta_0]\}, \tag{3.14}$$

which allows us to investigate the properties of T_{TW} by conditioning on the sufficient statistics S_t in (3.2) when (t, S_t) is on the boundary of the stopping region of $T_{\mathrm{M}}(c_0)$.

3.3.2 Asymptotic Properties

In this subsection, we present the asymptotic properties of our proposed tandem-width sequential CIs defined by the stopping time T_{TW} in (3.13). The main theoretical challenge is to investigate the asymptotic properties of the stopping time $T_{\mathrm{M}}(c)$ in (3.10) for fixed-width sequential CI as $h \to 0$, or equivalently, as $c = c_h \to 0$, including both asymptotic expression of ARL and the asymptotic CP. It is useful to point out that our method is applicable to derive the asymptotic properties of Frey's stopping time $T_{\mathrm{F}}(a, h)$ in (3.9), which has been developed in the literature, as Frey [23] only reports finite-sample numerical properties.

Let us begin with the investigation of the asymptotic properties of the stopping times $T_{\mathrm{M}}(c)$ in (3.10). We first investigate the asymptotic properties of $T_{\mathrm{M}}(c)$ including CP in the unconstraint scenario as the threshold $c \to 0$. Later this will allow us to investigate the constraint scenario by finding c that satisfy the CP constraint in (3.7).

The following theorem summarizes the main results for our proposed stopping times $T_{\mathrm{M}}(c)$ in (3.10) for fixed-width sequential CI.

Theorem 3.1. *As $c \to 0$, we have $T_{\mathrm{M}}(c) \to \theta(1-\theta)/c$ almost surely for each $\theta \in (0, 1)$,* *and*

$$\mathsf{E}_\theta[T_{\mathrm{M}}(c)] = (1+o(1))\frac{\theta(1-\theta)}{c}. \tag{3.15}$$

Moreover, denote by $\hat{\theta}_{T_{\mathrm{M}}}$ the MLE of θ in (3.2) at time $T_{\mathrm{M}}(c)$. Then, as $c \to 0$,

$$\frac{1}{\sqrt{c}}(\hat{\theta}_{T_{\mathrm{M}}} - \theta) \to N(0,1) \textit{ in distribution} \tag{3.16}$$

and thus an asymptotic $100(1-\alpha)\%$ CI for θ is $\hat{\theta}_{T_{\mathrm{M}}} \pm z_{\alpha/2}\sqrt{c}$.

Before detailing the proof of this theorem, it is useful to comment on the usefulness of the theorem. First, in $T_{\mathrm{M}}(c)$, if we set the half-width of the asymptotic $100(1-\alpha)\%$ CI for θ to be h, then $z_{\alpha/2}\sqrt{c} = h$ and thus $c = (h/z_{\alpha/2})^2$. This justifies the form of $c = c_h$ in (3.11) and shows that $\gamma \sim \alpha$ as $h \to 0$. Moreover, for $T_{\mathrm{M}}(c_0)$, with the threshold $c_0 = \rho_0 c$ for some constant $\rho_0 > 0$, as $c_0 \to 0$, we have $\mathsf{P}_\theta(\hat{\theta}_{T_{\mathrm{M}}(c_0)} \in [\theta_0, 1-\theta_0])$ is equal to $1 - o(1)$ if $\theta \in [\theta_0, 1-\theta_0]$ and $o(1)$ otherwise. When the sample sizes of these two stages are of the same order, then the $o(1)$ term will become negligible. Thus, for our proposed stopping time for tandem-width CI, the asymptotic properties follow directly from the the theorem if the thresholds c_0 and c_1 in the two stages are of the same order. Such results can be summarized in the following corollary.

Corollary 3.1. *For the proposed stopping time $T_{\mathrm{TW}}(c)$ in* (3.13) *with T_0 and T_1 being the stopping time $T_{\mathrm{M}}(c)$ in* (3.10) *with the thresholds $c_0 = \rho_0 c$ and $c_1 = \rho_1 c$, respectively, for some constant $\rho_0 > \rho_1 > 0$. Then as $c \to 0$, with probability 1 under P_θ, we have*

$$T_{\mathrm{TW}}(c) = \begin{cases} T_0, & \textit{if } \theta \in [\theta_0, 1-\theta_0]; \\ T_1, & \textit{if } \theta < \theta_0 \textit{ or if } \theta > 1-\theta_0. \end{cases} \tag{3.17}$$

and

$$
\mathsf{E}_\theta[T_{\mathrm{TW}}(c)] = \begin{cases} (1+o(1))\dfrac{\theta(1-\theta)}{c_0}, & \textit{if } \theta_0 < \theta < 1-\theta_0; \\[2ex] (1+o(1))\dfrac{\theta(1-\theta)}{c_1}, & \textit{if } \theta < \theta_0 \textit{ or } \theta > 1-\theta_0; \\[2ex] (1+o(1))\theta(1-\theta)(\dfrac{1}{2c_0}+\dfrac{1}{2c_1}), & \textit{if } \theta = \theta_0 \textit{ or } \theta = 1-\theta_0. \end{cases} \tag{3.18}
$$

Let us now prove Theorem 3.1. Before presenting the proof, it is useful to present two simple lemmas: one shows that $T_{\mathrm{M}}(c)$ in Theorem 3.1 is bounded above, and the other shows that $T_{\mathrm{M}}(c)$ is bounded below. Both bounds are non-asymptotic and hold for any threshold $c > 0$.

Lemma 3.1. *For $T_{\mathrm{M}}(c)$ in Theorem 3.1, we have $T_{\mathrm{M}}(c) \leq \max(1, 1/(4c))$ for any $c > 0$.*

Proof. The key idea is to note that $\theta_t^\star(1-\theta_t^\star) \leq 1/4$ regardless of the values of $\theta_t^\star$. When $t > 1/(4c) \geq 1$, we have

$$
\frac{\theta_t^\star(1-\theta_t^\star)}{t} \leq \frac{1}{4t} < \frac{1}{4 \cdot (1/(4c))} = c.
$$

The lemma then follows directly from the definition of the stopping time in (3.10). □

Lemma 3.2. *For $T_{\mathrm{M}}(c)$ in Theorem 3.1, we have $T_{\mathrm{M}}(c) \geq \left(\frac{1}{8c}\right)^{2/3}$ for any $c > 0$.*

Proof. By the definition of the minimax estimator $\theta_t^\star$ in (3.4), an elementary argument shows that for all $t \geq 1$,

$$
\frac{\theta_t^\star(1-\theta_t^\star)}{t} = \frac{S_t(t-S_t) + t\sqrt{t}/2 + t/4}{t(t+\sqrt{t})^2} \geq \frac{0 + t\sqrt{t}/2 + 0}{t(t+\sqrt{t})^2} > \frac{t\sqrt{t}/2}{t(t+t)^2} = \frac{1}{8}t^{-3/2}.
$$

Here we use the fact that $S_t(t-S_t) \geq 0$ since $0 \leq S_t = \sum_{i=1}^t X_i \leq t$. Hence, whenever $t \leq \left(\frac{1}{8c}\right)^{2/3}$, we have $\frac{\theta_t^\star(1-\theta_t^\star)}{t} > c$, and thus $T_{\mathrm{M}}(c)$ will not stop at time t. This proves the lemma. $\square$

Remark 3.1. Lemmas 3.1 and 3.2 provide non-asymptotic bounds that allow us to prove the asymptotic results in Theorem 3.1 as $c \to 0$ for our stopping time $T_{\mathrm{M}}(c)$. However, these results also apply to Frey's procedure $T_{\mathrm{F}}(c, a)$ in (3.9). In particular, by the elementary arguments in Lemmas 3.1 and 3.2 we can show that for $a > 0$ and $c > 0$

$$\sqrt{a/c} - 2a \leq T_{\mathrm{F}}(c, a) \leq \max(1, 1/(4c)), \tag{3.19}$$

which results in $T_{\mathrm{F}}(c, a) \to \infty$ almost surely as $c \to 0$.

Given the non-asymptotic bounds in Lemmas 3.1 and 3.2, we are now ready to prove the asymptotic results in Theorem 3.1 as $c \to 0$.

Proof of Theorem 3.1: By Lemma 3.2, as $c \to 0$, we have $T_{\mathrm{M}}(c) \to \infty$ with probability 1. To find an accurate asymptotic expression of $T_{\mathrm{M}}(c)$, it is useful to rewrite its stopping rule in terms of the MLE $\hat{\theta}_t$ in (3.4) whose asymptotic properties are well-known. A simple math argument shows that

$$\frac{\theta_t^{\star}(1 - \theta_t^{\star})}{t} = \frac{\theta(1-\theta) + 1/(2\sqrt{t}) + 1/(4t) + \Big(\hat{\theta}_t(1 - \hat{\theta}_t) - \theta(1-\theta)\Big)}{t(1 + 1/\sqrt{t})^2}. \tag{3.20}$$

At the high-level, the proof is based on two disjoint events of $\hat{\theta}_t$, depending on how close the term in (3.20) is to $\theta(1-\theta)/t$. By the law of large numbers, for any given $0 < \theta < 1$, the term in (3.20) is asymptotically equivalent to $\theta(1-\theta)/t$ with probability that tends to 1 for large t. This turns out to capture the first-order asymptotic analysis, as the corresponding complement event is negligible, since $T_{\mathrm{M}}(c)$ is bounded from above by Lemma 3.1.

Below is the detailed, rigorous proof. Fix $0 < \theta < 1$. Fix $\epsilon > 0$. Then, there exists an integer $t_\epsilon > 0$ such that for all $t \geq t_\epsilon$,

$$(1 + 1/\sqrt{t})^2 \leq 1 + \epsilon. \tag{3.21}$$

Furthermore, all $t \geq t_\epsilon$, denote the event

$$\mathcal{A}_{t,\epsilon} = \{|1/(2\sqrt{t}) + 1/(4t) + \Big(\hat{\theta}_t(1-\hat{\theta}_t) - \theta(1-\theta)\Big)| \leq \epsilon \cdot \theta(1-\theta)\}. \tag{3.22}$$

For the case when the event $\mathcal{A}_{t,\epsilon}$ does not hold, for $\epsilon > 0$ and for $\delta > 0$, by the weak law of large numbers, there exists $t_{\epsilon,\delta} > 0$ such that for $t \geq t_{\epsilon,\delta}$, $\mathsf{P}(\mathcal{A}^c_{t,\epsilon}) < \delta$. Moreover,

$$\mathsf{E}[T_{\mathrm{M}}] = \mathsf{E}[T_{\mathrm{M}}; \mathcal{A}_{t,\epsilon}] + \mathsf{E}[T_{\mathrm{M}}; \mathcal{A}^c_{t,\epsilon}]. \tag{3.23}$$

By Lemma 3.1, $T_{\mathrm{M}} \leq 1/(4c)$, so

$$\mathsf{E}[T_{\mathrm{M}}; \mathcal{A}_{t,\epsilon}] \leq \mathsf{E}[T_{\mathrm{M}}] \leq \mathsf{E}[T_{\mathrm{M}}; \mathcal{A}_{t,\epsilon}] + \frac{1}{4c}\mathsf{P}(\mathcal{A}^c_{t,\epsilon}) < \mathsf{E}[T_{\mathrm{M}}; \mathcal{A}_{t,\epsilon}] + \frac{1}{4c}\delta. \tag{3.24}$$

Now, we prove the case when the event $\mathcal{A}_{t,\epsilon}$ is true. In this case, a combination of (3.20) and given that the event $\mathcal{A}_{t,\epsilon}$ holds yields that for all $t \geq t_\epsilon$,

$$\frac{(1-\epsilon)\theta(1-\theta)}{(1+\epsilon)t} \leq \frac{\theta^\star_t(1-\theta^\star_t)}{t} \leq \frac{(1+\epsilon)\theta(1-\theta)}{t}. \tag{3.25}$$

Note that such t_ϵ might depend on θ and ϵ, but relation (3.25) holds for all $t \geq t_\epsilon$. Now by Lemma 3.2, there exists a $c^* > 0$ such that for all $c \leq c^*$, we have $T_{\mathrm{M}}(c) \geq t_\epsilon + 1$, and thus relation (3.25) holds to both $T_{\mathrm{M}}(c)$ and $T_{\mathrm{M}}(c) - 1$.

Now by the definition in (3.10), when $t = T_{\mathrm{M}}(c)$, we have $\frac{\theta^\star_t(1-\theta^\star_t)}{t} \leq c$. Combining this with the first inequality in (3.25) for $t = T_{\mathrm{M}}(c)$ yields that for all $c \leq c_\epsilon$,

$$\frac{(1-\epsilon)\theta(1-\theta)}{(1+\epsilon)T_{\mathrm{M}}(c)} \leq c, \qquad \text{or equivalently,} \quad cT_{\mathrm{M}}(c) \geq \frac{1-\epsilon}{1+\epsilon}\theta(1-\theta)$$

Letting $c \to 0$ yields that

$$\liminf_{c\to 0}\{cT_{\mathrm{M}}(c)\} \geq \frac{1-\epsilon}{1+\epsilon} \cdot \theta(1-\theta)$$

with probability $1-\delta$ for any given $\epsilon > 0$. Now the inf-limit in the left-hand side does not depend on ϵ. Letting $\epsilon \to 0$, we have, with probability $1-\delta$,

$$\lim_{c\to 0} \inf\{cT_{\mathrm{M}}\} \geq \theta(1-\theta). \tag{3.26}$$

On the other hand, by the definition in (3.10), when $t = T_{\mathrm{M}}(c)-1$, we have $\frac{\theta_t^\star(1-\theta_t^\star)}{t} > c$. Combining this with the second inequality in (3.25) for $t = T_{\mathrm{M}}(c)-1$ yields that for all $c \leq c_\epsilon$,

$$c < \frac{(1+\epsilon)\theta(1-\theta)}{T_{\mathrm{M}}(c)-1}, \qquad \text{or equivalently,} \quad cT_{\mathrm{M}}(c) < (1+\epsilon)\theta(1-\theta)+c$$

with probability $1-\delta$. Letting $c \to 0$, we have

$$\lim_{c\to 0} \sup\{cT_{\mathrm{M}}(c)\} \leq (1+\epsilon)\theta(1-\theta)$$

for any $\epsilon > 0$, which results in

$$\lim_{c\to 0} \sup\{cT_{\mathrm{M}}(c)\} \leq \theta(1-\theta). \tag{3.27}$$

with probability $1-\delta$. Combining (3.26) and (3.27),

$$\lim_{c\to 0}\{cT_{\mathrm{M}}(c)\} = \theta(1-\theta).$$

with probability $1-\delta$. By Lemma 3.1 and Lebesgue's dominated convergence theorem, we have

$$\lim_{c\to 0} \mathsf{E}_\theta[cT_{\mathrm{M}}; \mathcal{A}_{t,\epsilon}] = \mathsf{E}_\theta[\lim_{c\to 0} cT_{\mathrm{M}}; \mathcal{A}_{t,\epsilon}] = \theta(1-\theta),$$

and so

$$\mathsf{E}_\theta[T_{\mathrm{M}}(c); \mathcal{A}_{t,\epsilon}] = (1+o(1))\frac{\theta(1-\theta)}{c}. \tag{3.28}$$

Now, letting $\delta \to 0$ and using (3.28),

$$(1+o(1))\frac{\theta(1-\theta)}{c} \le \mathsf{E}[T_{\mathrm{M}}] \le (1+o(1))\frac{\theta(1-\theta)}{c} + o(\frac{1}{c}) = (1+o(1))\frac{\theta(1-\theta)}{c}. \quad (3.29)$$

This proves (3.15).

To prove (3.16), a crucial step is to define a integer-valued constant $m = m_c = [\theta(1-\theta)/c]$ as $c \to 0$. On the one hand, by the central limit theorem, $\sqrt{m}(\hat{\theta}_m - \theta)/\sqrt{\theta(1-\theta)} = (\hat{\theta}_m - \theta)/\sqrt{c}$ is asymptotically normally $N(0,1)$ distributed, as $c \to 0$. On the other hand, $T = T_{\mathrm{M}}(c)$, we just showed that $T/m \to 1$ almost surely. By equation (2.43) in Theorem 2.40 of Siegmund [33, p. 23], we have

$$\sqrt{m}(\hat{\theta}_T - \hat{\theta}_m) \to 0 \text{ in probability.} \quad (3.30)$$

Combine these two results together yields (3.16), completing the proof of the theorem. □

3.3.3 Finite-Sample Numerical Computation

In this subsection, we discuss the numerical computation of the finite-sample properties of our proposed stopping times $T_{\mathrm{M}}(c)$ in (3.10) and $T_{\mathrm{TW}}(c)$ in (3.18), including the ARL, $\mathsf{E}_\theta(T)$, and the CP, $\mathsf{P}_\theta(|\hat{\theta}_T - \theta| \le h)$ at each θ. This will allow us to validate the asymptotic properties of our stopping times in the previous subsection as well as compare properties of different methods.

For a given stopping time T and its corresponding sequential CI, there are two approaches to compute its finite-sample properties, $\mathsf{E}_\theta(T)$ and $\mathsf{P}_\theta(|\hat{\theta}_T - \theta| \le h)$ for all $0 < \theta < 1$. The first one is an approximate Monte Carlo method based on repeated random sampling of Bernoulli(θ) random variables for each $0 < \theta < 1$. It is straightforward to implement such Monte Carlo method, although it is very time consuming to obtain accurate estimates of the ARL or CP properties over the whole interval $\theta \in (0,1)$, especially when the true θ is close to 0 or 1. The second approach is an accurate non-Monte-Carlo numerical method based on the path-counting ideas in Girshick, Mosteller, and Savage [34] and Schultz, Nichol, Elfring, and Weed [35], also see Frey [23]. This non-Monte-Carlo numerical method is validated against the Monte Carlo method, and both yield the same results. Note that for schemes that satisfy the assumption of bounded stopping time such as $T_{\mathrm{M}}(c)$, the corresponding performance can also be computed numerically by applying the recursions of Lemma 2.4 in Chapter 2 without the need to perform Monte-Carlo simulations.

Below let us provide a more detailed discussion on the accurate non-Monte-Carlo numerical method. Note that $S_t = \sum_{i=1}^{t} X_i = S_{t-1} + X_t$ is a sufficient statistic for the Bernoulli proportion θ, and conditional on S_{t-1}, the value of S_t has only two choices: S_{t-1} or $S_{t-1}+1$, depending on whether $X_t = 0$ or 1. Then the key idea of the non-Monte-Carlo numerical method is to count the number of paths, denoted by $H(a,t)$, from $S_0 = 0$ at time 0 to $S_t = a$ at time t without hitting any earlier stopping boundaries of T before time t. For many reasonable stopping times T including our proposed stopping time $T = T_{\mathrm{M}}(h)$, the stopping points/bondaries of T can be written as the set of discrete points, $(S_{t_1} = a_1, t_1), \cdots, (S_{t_k} = a_k, t_k)$, for some (possibly large) $k \geq 1$. Also in our proposed stopping time and many other stopping times, $(S_{t_i} = a, t_i)$ is a stopping point if and only if $(t_i - a, t_i)$ is also a stopping point, due to the fact that the problem is symmetric at $p = 1/2$. Then when the stopping time T stops at time t_i with the observed value $S_{t_i} = a_i$, i.e., when (a_i, t_i) is a stopping point, we estimate θ by $\hat{\theta}_T = \hat{\theta}_i = a_i/t_i$ and report the confidence interval as $[\max(0, \hat{\theta}_T - h), \min(1, \hat{\theta}_T + h)]$.

Now once we have counted the number $H(a_i, t_i)$ of sample paths from $(0,0)$ to $(t = t_i, S_t = a_i)$ without hitting any earlier stopping regions for all stopping points of T, we can compute the finite-sample properties of T simultaneously for all θ by

$$\mathsf{P}_\theta(|\hat{\theta}_T - \theta| \leq h) = \sum_{i=1}^{k} H(a_i, t_i)\theta^{a_i}(1-\theta)^{t_i - a_i}\mathbb{1}\{|\theta - \hat{\theta}_i| \leq h\}, \text{and} \tag{3.31}$$

$$\mathsf{E}_\theta(T) = \sum_{i=1}^{k} H(a_i, t_i)\theta^{a_i}(1-\theta)^{t_i - a_i} t_i. \tag{3.32}$$

Numerically, we can use (3.31) and (3.32) to compute $\mathsf{P}_\theta(|\hat{\theta}_T - \theta| \leq h)$ and $\mathsf{E}_\theta(T)$ as a function θ as θ varies from 0 to 1 (or to $1/2$ due to symmetric properties) with a small step size.

For each threshold c or tuning parameter γ in (3.11), we will be able to derive the corresponding finite-sample properties, $\mathsf{E}_\theta(T)$ and $\mathsf{P}_\theta(|\hat{\theta}_T - \theta| \leq h)$, of our proposed stopping times $T = T_{\mathrm{M}}(c)$ in (3.10) or $T = T_{\mathrm{TW}}(c)$ in (3.18) for all $0 < \theta < 1$. To satisfy the $1-\alpha$ CP constraints in (3.1) or (3.7), we propose to use the bisection search method to obtain the desired threshold c or γ.

We split the remainder of this subsection into two parts: (a) the numerical computation of the finite-sample properties of $T = T_{\mathrm{M}}(c)$ in (3.10) and (b) the numerical computation of the finite-sample properties of $T = T_{\mathrm{TW}}(c)$ in (3.18). The latter part uses the numerical computations of part (a) but is more involved in computation because the stopping region for the tandem method involves two stopping regions, one from the first stage using h_0 and another from the second stage using h_1.

Finite-Sample Properties of $T_{\mathrm{M}}(c)$

Let us first focus on how to count the number of paths for a stopping time T such as $T = T_{\mathrm{M}}(c)$ in (3.10) whose stopping region boundary is *convex*. Without loss of generality, assume that the stopping time is defined as $T = \inf\{t \geq 1 : S_t \in \mathcal{R}_t\}$, where $\mathcal{R}_t = \mathcal{R}_t(\gamma)$ is the stopping region at time t. Note that $0 \leq S_t \leq t$ for all $t \geq 1$. Now for each t and each possible value $S_t = a$, we define two functions: one is the indicator function $I(a,t) = 1$ if $S_t = a$ is an interior (non-stopping) at time t and $I(a,t) = 0$ if $S_t = a$ belongs to the stopping region $\mathcal{R}_t$, and the other is the counting function $H(a,t)$ that denotes the number of ways to get to $S_t = a$ successes at time step t without hitting any earlier stopping regions $\mathcal{R}_k$'s at time $1 \leq k \leq t-1$. Note that $H(0,1) = H(1,1) = 1$, since we only have one way

Table 3.2: Choices of $\gamma = \gamma(h, \alpha)$ and $c = c(h, \alpha)$ for 90%, 95%, and 99% CIs of fixed half-width h for our method.

	$1-\alpha = 90\%$		$1-\alpha = 95\%$		$1-\alpha = 99\%$	
h	γ	c	γ	c	γ	c
0.10	0.0736	3.1242×10^{-3}	0.0351	2.2521×10^{-3}	0.0051	1.2749×10^{-3}
0.09	0.0762	2.5762×10^{-3}	0.0373	1.8678×10^{-3}	0.0057	1.0598×10^{-3}
0.08	0.0801	2.0895×10^{-3}	0.0394	1.5082×10^{-3}	0.0064	8.6090×10^{-4}
0.07	0.0826	1.6263×10^{-3}	0.0412	1.1757×10^{-3}	0.0071	6.7610×10^{-4}
0.06	0.0851	1.2143×10^{-3}	0.0426	8.7566×10^{-4}	0.0078	5.0856×10^{-4}
0.05	0.0877	8.5731×10^{-4}	0.0436	6.1395×10^{-4}	0.0086	3.6211×10^{-4}
0.04	0.0901	5.5699×10^{-4}	0.0450	3.9815×10^{-4}	0.0089	2.3382×10^{-4}
0.03	0.0925	3.1799×10^{-4}	0.0462	2.2646×10^{-4}	0.0092	1.3267×10^{-4}
0.02	0.0950	1.4350×10^{-4}	0.0475	1.0184×10^{-4}	0.0095	5.9468×10^{-5}
0.01	0.0975	3.6417×10^{-5}	0.0488	2.5759×10^{-5}	0.0097	1.4950×10^{-5}

to obtain $S_1 = 0$ or 1 at time $t = 1$.

To compute the counting function $H(S_t = a, t)$ in general, note that $S_{t-1} = a$ or $a - 1$ if $S_t = a$, depending on whether $X_t = 1$ or 0, and thus the number of path counts for points $(S_t = a, t)$ can be computed by the number of paths to either $(S_{t-1} = a, t-1)$ or $(S_{t-1} - 1 = a - 1, t - 1)$, when at least one of them is an interior (non-stopping) point. In other words, the counting function $H(S_t = a, t)$ can be recursively computed by

$$H(a,t) = H(a, t-1)I(a, t-1) + H(a-1, t-1)I(a-1, t-1), \tag{3.33}$$

where $I(a, t-1)$ and $I(a-1, t-1)$ are the indicator functions from the definition of the stopping time whether $S_{t-1} = a$ or $a - 1$ are interior (non-stopping) points at time $t - 1$ or not. For the purpose of numerical computation, the value $H(a, t)$ can be large for large t, and in such case, this recursion can be implemented on the log scale to avoid overflow problems by using the equality $\log(c + d) = \log c + \log(1 + \exp(\log d - \log c))$.

Table 3.2 below presents the numerical values of γ for different choices of α and h that guarantee that the coverage probability of the confidence interval is at least $1 - \alpha$.

Finite-Sample Properties of $T_{\mathrm{TW}}(c)$

It is much more challenging to count the number of paths for the stopping $T_{\mathrm{TW}}(c)$ in (3.18) for tandem-width sequential CIs, since its stopping region boundary is *non-convex*.

To better illustrate the challenges, consider Fig. 3.1, that plots the stopping points for our proposed tandem method with $h_0 = 0.1, h_1 = 0.05, \gamma_0 = 0.0351, \gamma_1 = 0.0436$ and $\theta_0 = 0.15$. Equivalently, $c_0 = 2.2521 \times 10^{-3}$ and $c_1 = 2.5759 \times 10^{-5}$. The stopping points in red represent the stopping points for $T_{\mathrm{M}}(c_0)$ when $h_0 = 0.1$ and $\hat{\theta}_{T_{\mathrm{M}}(c_0)} \in [\theta_0, 1-\theta_0]$. This means if we hit these red stopping points, then we stop sampling and report the $100(1-\alpha)\%$ CI as $[\max(0, \hat{\theta}_{T_{\mathrm{M}}(c_0)} - h_0), \min(1, \hat{\theta}_{T_{\mathrm{M}}(c_0)} + h_0)]$. However, if we do not hit these points in the first stage and instead hit the green stopping points for $T_{\mathrm{M}}(c_0)$ where $\hat{\theta}_{T_{\mathrm{M}}(c_0)} \notin [\theta_0, 1-\theta_0]$, then we need to keep on sampling until we reach the blue stopping points for $T_{\mathrm{M}}(h_1)$ and report the $100(1-\alpha)\%$ CI as $[\max(0, \hat{\theta}_{T_{\mathrm{M}}(c_1)} - h_1), \min(1, \hat{\theta}_{T_{\mathrm{M}}(c_1)} + h_1)]$. As a result, the stopping region boundary of $T_{\mathrm{TW}}(c)$ in (3.18) consists of both red and blue stopping times, which form a non-convex set. The good news is that this non-convex set is the difference of two convex boundaries, which allows us to simplify the computations.

To be more concrete, we use the definition of our tandem stopping time in (3.14) to split the CP and ARL for the tandem procedure into three parts as follows. First, we compute CP and ARL achieved by using $T_{\mathrm{M}}(c_1)$, hitting the blue stopping points when the blue region is the only stopping region. Second, compute CP and ARL achieved by hitting the red stopping points, i.e. the stopping points for the first stage where we stop sampling using the equations from the finite-sample properties of $T_{\mathrm{M}}(c)$ subsection. The third part is the more demanding part, as we need to compute the number of ways to hit the blue stopping points starting from the red stopping points without hitting any stopping points in the process. We start the recursion (3.33) from each red stopping point as the origin and continue recursively until we hit the blue region. Then, we finish the third step by computing CP and ARL as in (3.31) and (3.32) but with the modified number of ways reaching these blue points. The CP and ARL for the tandem procedure can be combined

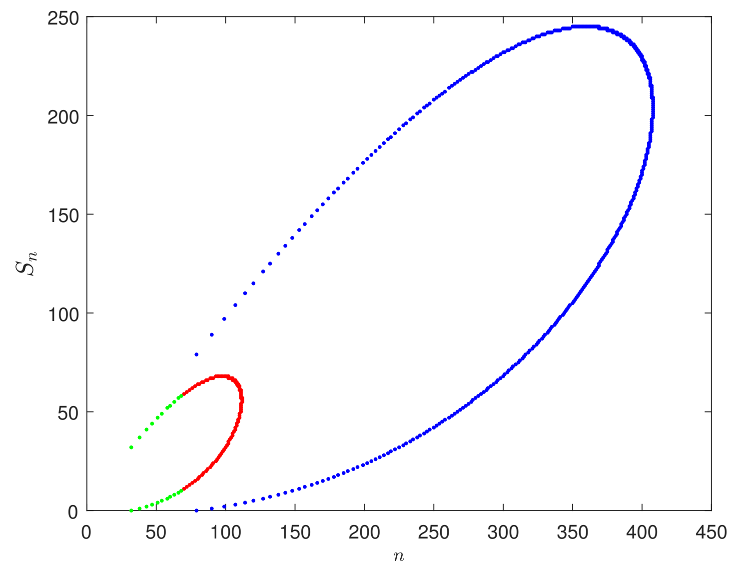


Figure 3.1: The stopping points for $T = T_{\text{TW}}$ in (3.14) with $h_0 = 0.1, h_1 = 0.05, \gamma_0 = 0.0351, \gamma_1 = 0.0436$ and $\theta_0 = 0.15$.

by adding the CP and ARL from the first and third parts and subtracting the second part.

3.4 Numerical and Simulation Examples

In this section, we report the numerical and simulation study results to further demonstrate the usefulness of our proposed stopping times. In Subsection 3.4.1, we illustrate the performance of the tandem-width stopping time T_{TW} in (3.13). In Subsection 3.4.2, we compare our proposed fixed-width stopping time T_{M} in (3.10) with Frey's method T_{F} in (3.9) that involves an additional tuning parameter of the Bayes prior.

3.4.1 Tandem-width CI

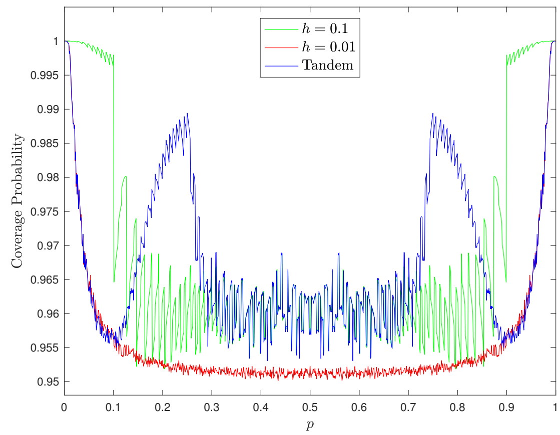
Suppose that we are interested in deriving a 95% tandem-width sequential CI with half-width $h_0 = 0.1$ if $\hat{p} \in [1 - \theta_0, \theta_0]$ for $\theta_0 = 0.15$ and with half-width $h_1 = 0.01$ if $\hat{\theta} < \theta_0 = 0.15$ or $> 1 - \theta_0 = 0.85$. For our proposed tandem-width CI method, two threshold values are $c_0 = 2.2521 \times 10^{-3}$ and $c_1 = 2.5759 \times 10^{-5}$, or equivalently, $\gamma_0 = 0.0351$ and $\gamma_1 = 0.0488$ based on Table 3.2. Next, we obtain the the coverage probability and ARL through simulation, with 500,000 replications at each value of $\theta = 0.001, 0.002, \ldots, 0.999$. Note that we could also use the path-counting ideas in the previous section to obtain CP and ARL analytically, but in this case it is easier to verify our results through simulation. We report the estimate of $\mathsf{P}_\theta(|\hat{\theta}_{T_{\mathrm{TW}}} - \theta| \leq h)$ as the number of instances that θ is within the reported confidence interval divided by the total number of replications. Furthermore, we report the estimate of $\mathsf{E}_\theta[T_{\mathrm{TW}}]$ as the average number of run lengths at each replication for each value of θ. In Figure 3.2, we compare the tandem-width CI simulation results (blue line) versus the analytical results (obtained through through the finite-sample numerical computational methods in Subsection 3.3.3) of the fixed-width CI based on $\theta_t^\star$ obtained with $h = 0.1$ (green line) and $h = 0.01$ (red line).

We can notice that by not choosing to use a fixed-width CI of $h = 0.01$, as that based on T_{M} in in (3.10), we can save in the worst case about 60 % of the sampling cost and time if we are willing to report a $100(1-\alpha)\%$ CI for θ with larger half-width $h = 0.1$ when θ is not close to 0 or close to 1. This saving in sampling cost becomes more obvious as we get closer to $p = 0.5$. This illustrates the importance of our tandem-width methodology, because when resources are scarce or when no historical data is available to gain prior knowledge about θ, then we do not need to spend so much time to report a very accurate CI with a very small half-width when θ is close to $1/2$.

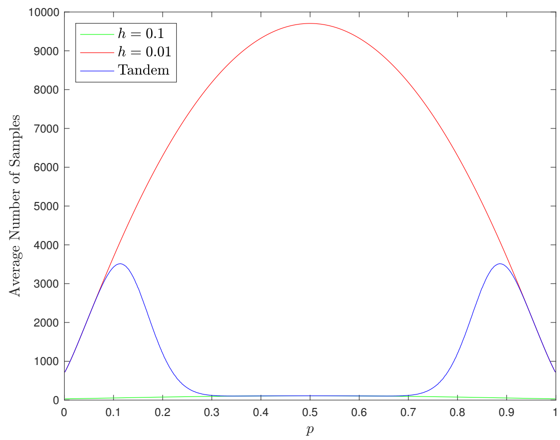
Now that we illustrated the importance of our tandem-width methodology, we compare the performance of the minimax-based method in (3.10) versus Frey's method in (3.9).

3.4.2 Fixed-width CI Comparisons

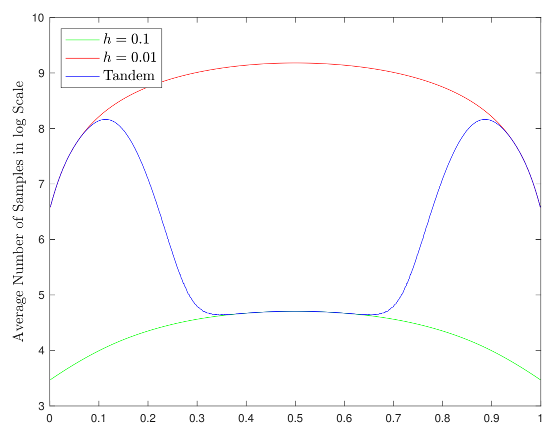
In this subsection, we compare our proposed fixed-width method with Frey's method T_{F} in (3.9) and with the optimum scheme in our earlier work in [29] and in Chapter 2. Using the numerical iterations from Subsection 3.3.3, we calculate numerically $\mathsf{P}_\theta(|\hat{\theta}_T - \theta| \leq h)$ and $\mathsf{E}_\theta(T)$ for $\theta = 1/2001, 2/2001, \ldots, 2000/2001$. Note that the requirement is to be able to guarantee a minimal coverage probability *for all* θ. Therefore, parameters were selected



(a) Coverage Probability



(b) Expected Sample Size



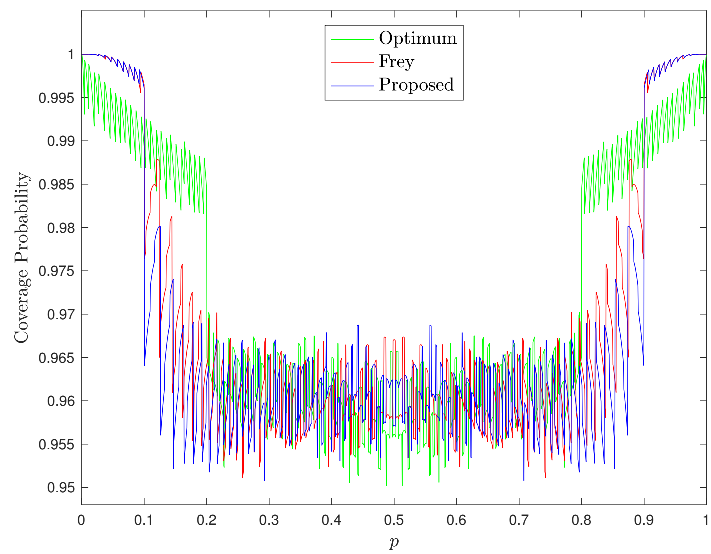
(c) Expected Sample Size in log Scale

Figure 3.2: A comparison of coverage probability and average run length for three sequential methods: (i) our proposed tandem-width CI with $h_0 = 0.01$ and $h_1 = 0.1$ (blue line); (ii) our proposed fixed-width CI with $h = h_0 = 0.01$ (red line); and (iii) our proposed fixed-width CI with $h = h_1 = 0.1$ (green line).

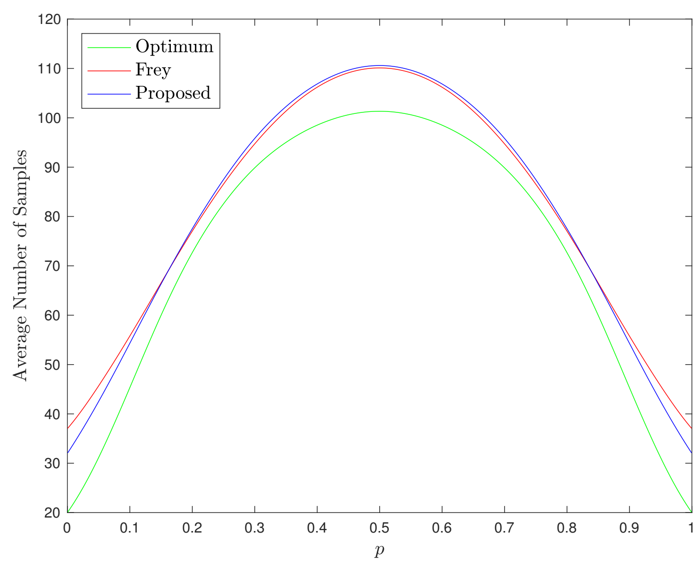
so that all competing schemes guaranteed *the same worst-case coverage probability*, i.e., coverage of at least $1 - \alpha$ for all θ. Here the tuning parameter γ is chosen from Table 3.2 for our method and from Table 3.1 for Frey's method T_{F}. The optimum scheme in [29] requires two tuning parameters: one is the parameter u that sets the Beta(u, u) as the prior distribution of θ, and the other is the parameter κ for the cost per observation. Here, for the choice of $u = 1$ (uniform prior), cost $\kappa = 0.00097$ will satisfy the coverage probability constraint in [29] for $\alpha = 0.05$ and $h = 0.1$.

In Fig. 3.3a, we plot the coverage probability for each test versus θ and in Fig. 3.3b, the corresponding average sample size required to obtain this performance for $\alpha = 0.05$ and $h = 0.1$. We can draw the following conclusions from the figures: Our proposed scheme and Frey's require about the same sample sizes for most values of θ, although our fixed-width scheme is slightly more parsimonious when θ is close to 0 or 1. Moreover, the two procedures exhibit similar coverage probability profiles. The optimum scheme in [29] is the best in terms of the smallest number of samples to guarantee the worst-case CP of at least 0.95.

We also ran numerical experiments for many other combinations of (α, h), and we make similar conclusions. For instance, in Fig. 3.4a, we plot the coverage probability for each test versus θ and in Fig. 3.4b, the corresponding average sample size required to obtain this performance for $\alpha = 0.05$ and $h = 0.05$. Our proposed method and Frey's method perform almost identically, whereas the optimum method has smaller sample size and larger coverage probability if the true θ is not too close to 0 or 1. Notice that the behavior of the optimum scheme differs between different values of h. For instance, for $h = 0.1$, the optimum scheme has a lower expected sample size than both methods, whereas for the case of $h = 0.05$ the expected sample size of the optimum scheme is sometimes larger than both methods, even though in such cases the coverage probability is larger. One possible explanation to this phenomenon is that the optimum scheme puts more weights on the expected sample size when h is larger, but more weights on CP when h is smaller. However, we are unable to prove such claim.



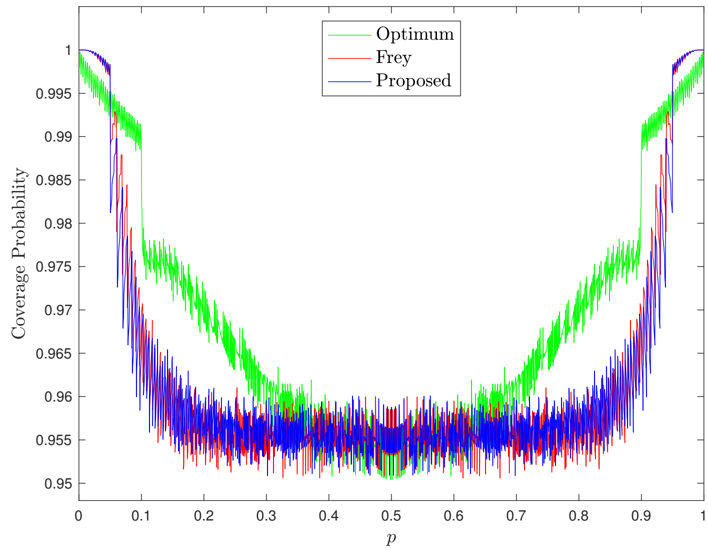
(a) Coverage Probability



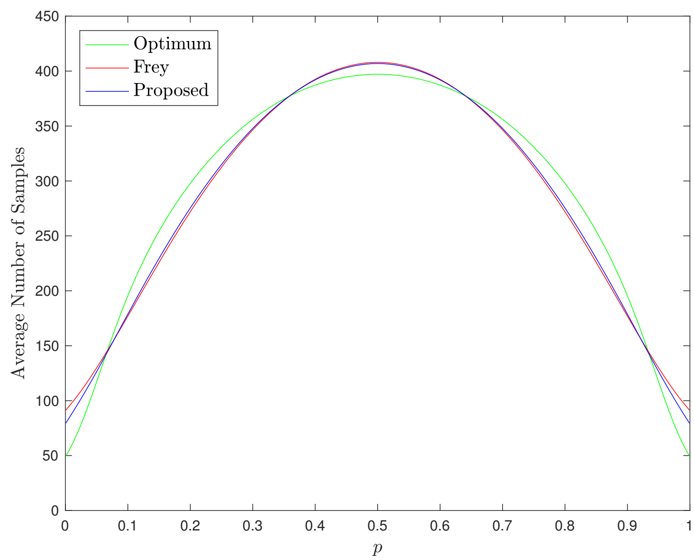
(b) Expected Sample Size

Figure 3.3: A comparison of coverage probabilities and expected sample sizes of three methods, for $h = 0.1$.

We should emphasize that the optimum scheme in our earlier work in [29] becomes computationally expensive as h gets smaller, e.g., $h = 0.01$, as it involves dynamic programming and the involvement of matrices of dimension of order $1/h^2$; see [29]. For the



(a) Coverage Probability



(b) Expected Sample Size

Figure 3.4: A comparison of coverage probabilities and expected sample sizes of three methods, for $h = 0.05$.

fixed half-width $h = 0.01$, the performances between our fixed-width method and Frey's are also similar, although Frey's method Frey [23] gives a slightly smaller (i.e. better) ARL whereas our proposed method gives a slightly larger coverage probability.

In summary, as compared to Frey's method Frey [23] that needs to optimize the tuning parameter for Bayes prior, our proposed method has similar finite-sample properties, but is much simpler since the minimax estimator does not involve any tuning parameters. In other words, our research on tandem-width sequential CI shed new light to develop simple but useful fixed-width sequential CI that is fast and efficient with performance characteristics that are comparable to or only slightly worse than those of the optimum scheme.

3.5 Conclusions

We proposed two sequential scheme for obtaining confidence intervals for a binomial proportion θ using the minimax estimator of θ: a fixed-width scheme, and a tandem-width scheme. We also established upper and lower bounds for our stopping times, presented some asymptotic properties, and compared the tandem-width procedure with the proposed fixed-width alternative. We also compared the performance of our fixed-width procedure with two existing sequential alternatives.

The main advantage of our proposed tandem-width method is that it can be extended to more than two stages or half-widths, allowing better flexibility with the choice of half-widths based on the true value of θ. For instance, one may prefer a half-width size of $h = 0.10$ if the true $\theta \in [0.4, 0.6]$, but may prefer a smaller half-width $h = 0.05$ if $\theta \in (0.1, 0.4)$ or $\theta \in (0.6, 0.9)$, or even a smaller half-width $h = 0.01$ if $\theta < 0.1$ or $\theta > 0.9$. Some existing literature formulate the sequential problem as a relative-width CI, where the half-width h is a function of θ, e.g. $h = h(\theta) = \eta\theta$ for some $\eta \in (0, 1)$. However, for small θ, say $\theta < 0.1$, this scheme becomes very costly especially with a small η. Our method takes care of this issue by fixing the half width for the interval when θ is small or large, avoiding this sampling cost concern.

Acknowledgements

This research was supported in part by NSF CMMI-1362876, through Georgia Institute of Technology, and in part by NSF CIF-1513373, through Rutgers University.

CHAPTER 4

SEQUENTIAL ESTIMATION BASED ON CONDITIONAL COST

4.1 Introduction

Extending the optimum methodology that is presented in Chapter 2 for Bernoulli trials to arbitrary distributions is generally analytically intractable and presents computational difficulties. In this chapter, we propose an alternative formulation to the optimum formulation that circumvents these analytical/computational difficulties for interval estimation for an arbitrary parameter of interest from an arbitrary distribution.

Parameter estimation is needed in numerous problems across different scientific fields. In most applications, estimation is primarily based on fixed-sample-size methodology. However, when we are interested in obtaining a reliable estimate as quickly as possible, then it is necessary to resort to sequential techniques. It is well known that in hypothesis testing, sequential methods [36] often enjoy significant reduction in the number of samples required to reach a reliable decision as compared to fixed-sample-size alternatives. Therefore, it is only natural to expect that this important advantage will carry over to estimation as well. Before addressing the problem of sequential estimation, let us first introduce some necessary background knowledge regarding classical estimation.

We observe a collection of random variables $\{X_1, \ldots, X_t\}$, where $t > 0$ is an integer. For simplicity, we assume $\{X_t\}$ is independent and identically distributed (i.i.d.) with a common probability density function (pdf) $\mathsf{f}(x|\theta)$ and parameter θ is considered random with a known prior pdf $\pi(\theta)$. Regarding the process $\{X_t\}$, the samples are generated as follows: Nature randomly selects the parameter θ following $\pi(\theta)$; then keeping θ fixed, Nature generates the sequence $\{X_t\}$ following $\mathsf{f}(x|\theta)$. It is therefore clear that the joint pdf of the set of samples $\{X_1, \ldots, X_t\}$ and θ has the following form

$$\mathsf{f}_t(x_1, \ldots, x_t, \theta) = \pi(\theta) \cdot \mathsf{f}(x_1|\theta) \cdots \mathsf{f}(x_t|\theta). \tag{4.1}$$

The joint pdf induces a probability measure which we denote by $\mathsf{P}(\cdot)$ while we reserve the symbol $\mathsf{E}[\cdot]$ for the corresponding expectation. If we also denote with $\mathscr{F}_t = \sigma\{X_1, \ldots, X_t\}$ the sigma-algebra generated by the first t samples, then we can write the conditional (posterior) pdf of θ given $\mathscr{F}_t$ as

$$\pi_t(\theta|\mathscr{F}_t) = \frac{\pi(\theta) \prod_{j=1}^{t} \mathsf{f}(x_j|\theta)}{\int \pi(\theta) \prod_{j=1}^{t} \mathsf{f}(x_j|\theta)\, d\theta}. \tag{4.2}$$

Equations (4.1) and (4.2) describe completely the statistical behavior of our observations. The goal is, using the acquired samples, to estimate the specific realization of θ that generates the data.

When we have a fixed sample size $\{X_1, \ldots, X_t\}$, then the problem of optimum estimation is solved very efficiently by following the Bayesian formulation [37, pp. 142–156]. Specifically let $\hat{\theta}(X_1, \ldots, X_t)$ denote any functions of the observations which can serve as a potential estimator of θ. Assume we are given a cost function $\mathsf{C}(\hat{\theta}, \theta)$ and consider the average cost $\mathsf{E}[\mathsf{C}(\hat{\theta}, \theta)]$, where averaging is with respect to the observations and θ. We are interested in finding the estimator that minimizes this expression. In other words we would like to perform the minimization $\inf_{\hat{\theta}} \mathsf{E}[\mathsf{C}(\hat{\theta}, \theta)]$ which leads to the classical Bayes estimator.

To find our estimator, we compute the *conditional* average cost

$$\mathsf{E}[\mathsf{C}(\hat{\theta}, \theta)|\mathscr{F}_t] = \int_{\mathbb{R}} \mathsf{C}(\hat{\theta}, \theta)\pi_t(\theta|\mathscr{F}_t)\, d\theta. \tag{4.3}$$

Then, it is well known that the optimum Bayes estimator satisfies

$$\hat{\vartheta}_t = \arg\inf_{\hat{\theta}} \mathsf{E}[\mathsf{C}(\hat{\theta}, \theta)|\mathscr{F}_t], \tag{4.4}$$

and the corresponding minimum conditional average cost is given by

$$\mathscr{C}_t = \inf_{\hat{\theta}} \mathsf{E}[\mathsf{C}(\hat{\theta}, \theta)|\mathscr{F}_t] = \mathsf{E}[\mathsf{C}(\hat{\vartheta}_t, \theta)|\mathscr{F}_t]. \tag{4.5}$$

Both $\hat{\vartheta}_t$ and $\mathscr{C}_t$ are $\mathscr{F}_t$-measurable since they are functions of the available observations.

4.2 Sequential Estimation

Under a sequential setup, process $\{X_t\}$ is acquired sequentially. At each time t we observe the accumulated information $\mathscr{F}_t$ which grows with time, thus generating the filtration $\{\mathscr{F}_t\}$ and the sequence $\{\mathsf{f}_t(\cdot)\}$ of joint pdfs. We use the same symbols $\mathsf{P}(\cdot)$ and $\mathsf{E}[\cdot]$ to denote the corresponding probability measure and expectation. One would be interested in defining a stopping time T that is adapted to $\{\mathscr{F}_t\}$ and a corresponding estimator $\hat{\theta}_T$ that is $\mathscr{F}_T$-measurable in order to provide an estimate of θ.

Since our goal is to limit the number of samples needed to compute the estimate, we would like to find a pair $(T, \hat{\theta}_T)$ that minimizes the average number of samples $\mathsf{E}[T]$ while, at the same time, we control the average estimation cost. To be more precise, we would like to consider the following constrained optimization problem for the determination of the optimum pair

$$\inf_{T, \hat{\theta}_T} \mathsf{E}[T], \quad \text{subject to: } \mathsf{E}[\mathsf{C}(\hat{\theta}_T, \theta)] \leq \beta, \tag{4.6}$$

where β is a level selected by the scientist. It has been pointed out in the literature [38–41] that solving (4.6) presents computational challenges, and this problem is by no means analytically tractable.

4.2.1 Alternative Optimization Problem

The analytical difficulties we mentioned can in fact be circumvented if we are willing to sacrifice part of our performance. We therefore propose to replace the constraint in (4.6) with the following *conditional* alternative

$$\mathsf{E}[\mathsf{C}(\hat{\theta}_T, \theta)|\mathscr{F}_T] \leq \tilde{\beta}.$$

If for example we select $\tilde{\beta} = \beta$ then the previous conditional version assures validity of the unconditional constraint in (4.6). The proposed modification in the constraint suggests a corresponding optimization problem

$$\inf_{T, \hat{\theta}_T} \mathsf{E}[T], \quad \text{subject to: } \mathsf{E}[\mathsf{C}(\hat{\theta}_T, \theta)|\mathscr{F}_T] \leq \tilde{\beta}, \tag{4.7}$$

as an alternative to the original one in (4.6). The formulation of the parameter estimation problem with (4.7) is along the same lines of the approaches adopted in [40, 41] for Gaussian processes. We should also mention that similar ideas were used for simultaneous detection and estimation for Gaussian [42] and conditionally Gaussian [43] data.

Remark 4.1. Before continuing with the analysis and solution of our optimization, let us discuss the differences between the two approaches depicted in (4.6) and (4.7). We observe that in the first we can have realizations of the observation sequence for which, at the time of stopping, the conditional average cost will satisfy $\mathsf{E}[\mathsf{C}(\hat{\theta}_T, \theta)|\mathscr{F}_T] > \beta$. Inequalities in the "wrong" direction tend to require smaller sample sizes, thus contributing towards the reduction of $\mathsf{E}[T]$. As we can see, in (4.7) such inequalities are not permitted since we force the conditional average cost to be always below $\tilde{\beta}$ *for every realization of the observations*. Therefore if we select $\tilde{\beta} = \beta$ we will end up with a scheme that satisfies the constraint in (4.6) in the *strict* sense. For this reason we need to increase $\tilde{\beta}$ slightly and select $\tilde{\beta} > \beta$ in order to achieve exact equality.

Remark 4.2. We should emphasize that even with a value of $\tilde{\beta}$ selected so as to satisfy the constraint in (4.6) with equality, the scheme we obtain by solving (4.7) is *not* the optimum for (4.6). The hope, however, is that the performance degradation by solving (4.7) instead of (4.6) will not be overly dramatic. In any case, as we mentioned, because of this performance sacrifice, our estimation problem simplifies considerably allowing for the development of an analytic solution.

The optimizations depicted in (4.6) and (4.7) require the definition of a pair $(T, \hat{\theta}_T)$. In the sequel, using proper analysis, we are going to design a candidate pair $(\mathcal{T}, \hat{\theta}_{\mathcal{T}})$ and then we will demonstrate that it is in fact the one that solves the optimization problem of interest, namely, the problem in (4.7). We begin the presentation of $(\mathcal{T}, \hat{\theta}_{\mathcal{T}})$ by first introducing our estimator.

4.2.2 Candidate Estimator

Let us fix the stopping time T and attempt to find the estimator $\hat{\theta}_T$ that minimizes the conditional average cost $\mathsf{E}[\mathsf{C}(\hat{\theta}_T, \theta)|\mathscr{F}_T]$. Assuming T stops almost surely (a.s.), we can write

$$\begin{aligned}
\mathsf{E}[\mathsf{C}(\hat{\theta}_T, \theta)|\mathscr{F}_T] &= \mathsf{E}\left[\sum_{t=0}^{\infty} \mathsf{C}(\hat{\theta}_t, \theta)\mathbb{1}\{T = t\}|\mathscr{F}_t\right] \\
&= \sum_{t=0}^{\infty} \mathsf{E}\left[\mathsf{C}(\hat{\theta}_t, \theta)|\mathscr{F}_t\right] \mathbb{1}\{T = t\} \\
&\geq \sum_{t=0}^{\infty} \inf_{\hat{\theta}} \mathsf{E}\left[\mathsf{C}(\hat{\theta}, \theta)|\mathscr{F}_t\right] \mathbb{1}\{T = t\} \\
&= \sum_{t=0}^{\infty} \mathscr{C}_t \mathbb{1}\{T = t\} = \mathscr{C}_T.
\end{aligned} \tag{4.8}$$

We note that the indicator function $\mathbb{1}\{T = t\}$ can be moved outside the conditional expectation because it is $\mathscr{F}_t$-measurable. Furthermore, using (4.4) and (4.5) for each deterministic value of t, we provide a lower bound on the conditional average cost with its minimum value $\mathscr{C}_t$. It is also clear that the inequality in (4.8) becomes an equality if we select $\hat{\theta}_t$ to be the optimum Bayes estimator $\hat{\vartheta}_t$.

This result suggests that when we stop at T if we apply the optimum Bayes estimator to the available data $\mathscr{F}_T$, then the conditional expected cost $\mathsf{E}[\mathsf{C}(\hat{\theta}_T, \theta)|\mathscr{F}_T]$ matches the lower bound $\mathscr{C}_T$. Consequently, for any stopping time T, we propose as a candidate estimator the Bayes estimator $\hat{\vartheta}_T$.

4.2.3 Candidate Stopping Time

Let us now turn to the definition of the candidate stopping time. As observations accumulate, at each time instant t we can compute the corresponding Bayes estimate $\hat{\vartheta}_t$ and the resulting minimum conditional average cost $\mathscr{C}_t$. The sequence $\{\mathscr{C}_t\}$ that is generated by these sequential computations can serve to define our candidate stopping time as follows,

$$\mathcal{T} = \inf\{t \geq 0 : \ \mathscr{C}_t \leq \tilde{\beta}\}. \tag{4.9}$$

In other words, we monitor the sequence of minimum conditional average costs and the *first time* the value of $\mathscr{C}_t$ falls below $\tilde{\beta}$ is the time we stop.

Combining the two results, it is clear that we propose the pair $(\mathcal{T}, \hat{\vartheta}_{\mathcal{T}})$ for stopping and parameter estimation. More precisely, we suggest to stop at $\mathcal{T}$ defined in (4.9) and use the data obtained up to the time of stopping to compute the Bayes estimate $\hat{\vartheta}_{\mathcal{T}}$. With the next theorem we show that this choice is optimum in the sense that it solves the constrained optimization problem defined in (4.7).

Theorem: *Consider any competing pair $(T, \hat{\theta}_T)$ which satisfies the constraint*

$$\mathsf{E}[\mathsf{C}(\hat{\theta}_T, \theta)|\mathscr{F}_T] \leq \tilde{\beta}.$$

Assuming that $\mathcal{T}$ and T stop a.s., then for each realization of our observations we have $\mathcal{T} \leq T$.

Proof. Since the pair $(T, \hat{\theta}_T)$ satisfies the constraint in (4.7), this means that

$$\tilde{\beta} \geq \mathsf{E}[\mathsf{C}(\hat{\theta}_T, \theta)|\mathscr{F}_T] \geq \mathscr{C}_T.$$

The first inequality is due to our assumption and the second is a consequence of (4.8) where we fix T and minimize over $\hat{\theta}_T$. We can thus conclude that $\mathscr{C}_T \leq \tilde{\beta}$. But this inequality immediately implies $\mathcal{T} \leq T$. Indeed this is the case because $\mathcal{T}$ is the *first* time instant for which $\mathscr{C}_t \leq \tilde{\beta}$. We have thus proved that for *each realization* any competing stopping time T will be no less than the candidate stopping time $\mathcal{T}$. Clearly this also implies that $\mathsf{E}[\mathcal{T}] \leq \mathsf{E}[T]$. This argument proves that the proposed pair is the one solving the constrained optimization problem depicted in (4.7). □

We point out that if the constraint is very mild, namely $\tilde{\beta}$ is overly large, then our method can lead to a trivial optimum pair $(T, \hat{\theta}_T)$. Indeed it is possible to stop at $\mathcal{T} = 0$ a.s. and simply use the prior to provide the necessary estimate. This can happen when $\mathscr{C}_0 \leq \tilde{\beta}$, namely,

$$\mathscr{C}_0 = \inf_{\hat{\theta}} \int \mathsf{C}(\hat{\theta}, \theta)\pi(\theta)\, d\theta \leq \tilde{\beta},$$

leading to the deterministic estimate

$$\hat{\vartheta}_0 = \arg \inf_{\hat{\theta}} \int \mathsf{C}(\hat{\theta}, \theta)\pi(\theta)\, d\theta.$$

Consequently, in order to avoid such a trivial outcome we select $\tilde{\beta} < \mathscr{C}_0$.

Remark 4.3. We emphasize that the desired problem to solve is (4.6). It is because of its analytical intractability that we resort to (4.7) which is possible to solve efficiently. When, however, we study the performance of the scheme produced by (4.7), we must test its behavior with respect to the constraint in (4.6) and *not* the conditional version adopted in (4.7). In this sense, even though the pair $(\mathcal{T}, \hat{\vartheta}_{\mathcal{T}})$ is "optimum," it should not come as a surprise if its performance, in some cases, turns out to be inferior to that of the fixed-sample-size estimator.

4.3 Optimizing Coverage Probability

A major goal in parameter estimation is, of course, the design of an estimator but also the selection of a sample size that can assure that the estimate is within a prescribed (confidence) interval around the correct value with some minimal guaranteed (coverage) probability. Specifically we would like to find a sample size T, fixed or random (stopping time), and an estimator $\hat{\theta}_T$ of θ assuring that $\mathsf{P}(|\hat{\theta}_T - \theta| \leq h) \geq 1 - \alpha$ for a specified $\alpha \in (0, 1)$. Parameter $h > 0$ denotes the desired half-width of the confidence interval (CI) and $1 - \alpha$ the minimal level of the coverage probability.

This problem can be effectively treated using the general framework we introduced in the previous section by selecting $\mathsf{C}(\hat{\theta}, \theta) = 1 - \mathbb{1}\{|\hat{\theta} - \theta| \leq h\}$ and $\beta = \alpha$. The conditional average cost function, using (4.3), can be written as

$$\mathsf{P}(|\hat{\theta}_t - \theta| > h | \mathscr{F}_t) = 1 - \int_{\hat{\theta}_t - h}^{\hat{\theta}_t + h} \pi_t(\theta | \mathscr{F}_t)\, d\theta,$$

where in the integration one should take into account the (essential) support of θ as it is dictated by the prior $\pi(\theta)$; for example, if $\theta \geq 0$ a.s., then the lower integration boundary must be replaced by $(\hat{\theta}_t - h)^+$. The Bayes estimator and the corresponding minimum conditional average cost are given by

$$\hat{\vartheta}_t = \arg \sup_{\hat{\theta}} \int_{\hat{\theta} - h}^{\hat{\theta} + h} \pi_t(\theta | \mathscr{F}_t)\, d\theta, \tag{4.10}$$

and

$$\mathscr{C}_t = 1 - \int_{\hat{\vartheta}_t - h}^{\hat{\vartheta}_t + h} \pi_t(\theta | \mathscr{F}_t)\, d\theta. \tag{4.11}$$

As we will have the chance to verify from the examples that follow, working directly with the coverage probability most often results in estimators and conditional costs that do not have analytic expressions and need to be computed numerically.

Next we present three classical parameter estimation examples where we compute their estimators and stopping times and compare their performances with fixed-sample-size methods and existing sequential techniques.

4.3.1 Mean of a Gaussian

Let us begin by considering the classical problem of estimating the unknown mean of a Gaussian random variable. Suppose our i.i.d. observations $\{X_t\}$ are $X_t \sim \mathcal{N}(\theta, \sigma^2)$ and for the prior of the mean we have $\theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$, where $\mu, \sigma^2, \sigma_\theta^2$ are known. The first step in our analysis consists of computing the posterior pdf $\pi_t(\theta|\mathscr{F}_t)$. It is a simple exercise to verify that

$$\pi_t(\theta|\mathscr{F}_t) = \mathcal{N}\left(\mu_t, \sigma_t^2\right)$$

where

$$\mu_t = \frac{\sigma_\theta^2 \sum_{j=1}^t X_j + \mu\sigma^2}{\sigma_\theta^2 t + \sigma^2}; \quad \sigma_t^2 = \frac{\sigma_\theta^2 \sigma^2}{\sigma_\theta^2 t + \sigma^2}.$$

From (4.10) and (4.11) the Bayesian estimator can be found as follows

$$\hat{\vartheta}_t = \arg\sup_{\hat{\theta}} \left\{\Phi\Big(\frac{h + \hat{\theta} - \mu_t}{\sigma_t}\Big) - \Phi\Big(\frac{-h + \hat{\theta} - \mu_t}{\sigma_t}\Big)\right\} = \mu_t,$$

where $\Phi(\cdot)$ denotes the standard Gaussian cumulative density function (cdf) and

$$\mathscr{C}_t = 2\Phi\Big(-\frac{h}{\sigma_t}\Big).$$

Since $\mathscr{C}_t$ is purely deterministic it is clear that the resulting stopping time $\mathcal{T}$ in (4.9) will be *deterministic* as well. Actually, for this case we can even solve the original optimization problem (4.6) and the resulting optimum stopping time is still deterministic [38].

Remark 4.4. With this simple example we realize that sequential estimation does not necessarily enjoy similar consequences as sequential hypothesis testing (in fact, this is the reason we included this case). We recall that in hypothesis testing when deciding between $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$, optimum sequential techniques require, on average, *four times fewer* samples than optimum fixed-sample-size tests [37, p. 109]. When, however, we estimate the mean of a Gaussian random variable, as we have seen, there is absolutely no gain. Fortunately, this conclusion is not universal and in the next two examples we will experience gains that are worth reporting.

4.3.2 Bernoulli Trials

In Chapters 2 and 3, we explored two new methodologies for interval estimation for a binomial proportion. In fact, methods that estimate *proportions* accompanied by confidence intervals are being used in many applications as polls; surveys; determination of fractions of people, animals or goods having certain traits/characteristics; etc. In these problems minimizing the number of samples that are necessary to assure estimates of a given quality is, clearly, of paramount importance. The simplest and most common model used to describe the corresponding data is Bernoulli binary sequences, which is also the model we adopt here.

In the literature there are various fixed sample size estimators [7, 44] addressing the question of proportion estimation, but we can also find sequential methods involving stopping times. In particular in [38] the optimization problem defined in (4.6) for this specific example is treated under an asymptotic regime, while in [23, 45] stopping rules are proposed and numerically compared without being supported by any form of optimality. Please refer to Section 1.4 in Chapter 1 for a more comprehensive literature on CI's for a binomial proportion.

Let us analyze this estimation problem using the methodology we introduced in the previous section. Consider an i.i.d. process $\{X_t\}$ with binary samples $X_t \in \{0,1\}$ and $\mathsf{P}(X_t = 1) = \theta \in [0,1]$. Probability θ is the parameter to be estimated for which we assume to have a symmetric prior Beta(θ, a, a) with known parameter $a > 0$. Due to the limits of θ we additionally need to assume that $0 < h < 0.5$.

If we call $S_t = X_1 + \cdots + X_t$ then the conditional pdf $\pi_t(\theta|\mathscr{F}_t)$ satisfies

$$\pi_t(\theta|\mathscr{F}_t) = \frac{\theta^{a+S_t-1}(1-\theta)^{a+t-S_t-1}}{B(a+S_t, a+t-S_t)} \tag{4.12}$$

which is Beta$(\theta, a+S_t, a+t-S_t)$ distributed. Here, $B(r,s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$. Since $0 \leq \theta \leq 1$ if we apply (4.10) we can write

$$\mathsf{E}[\mathsf{C}(\hat{\theta},\theta)|\mathscr{F}_t] = 1 - I_{\min\{1,\hat{\theta}_t+h\}}(a+S_t, a+t-S_t) + I_{\max(0,\hat{\theta}_t-h)}(a+S_t, a+t-S_t),$$

where $I_x(p,q)$ is the incomplete Beta function (see [28, Page 944]), which is the cdf of Beta(θ, p, q).

For $S_t = 1-a$ the previous expressions is minimized by $\hat{\vartheta}_t = h$ and for $S_t = a+t-1$ with $\hat{\vartheta}_t = 1-h$. For any other value of S_t finding the Bayes estimator requires the *numerical* solution of the equation

$$\hat{\vartheta}_t = \arg\left\{\hat{\theta}_t : \left(\frac{\hat{\theta}_t - h}{\hat{\theta}_t + h}\right)^{a+S_t-1} = \left(\frac{1-h-\hat{\theta}_t}{1+h-\hat{\theta}_t}\right)^{a+t-S_t-1}\right\}$$

with $h \leq \hat{\theta} \leq 1-h$.

Note that these results are also presented in Chapter 2. Furthermore, in Figure 2.3, we plot the average number of samples $\mathsf{E}[T]$ versus the coverage probability $\mathsf{P}(|\hat{\theta}_T - \theta| \leq h)$ when $a = 1$ and $h = 0.05$ for the optimum scheme that solves (4.6), the proposed conditional scheme, Frey's [23] sequential method, and the optimum fixed-sample-size method. As we can see, the proposed method outperforms the fixed sample size and the estimator in [23]. We also observe that, as the coverage probability approaches 1, we enjoy bigger gains in sample size. These gains in sample size are close to those of the optimum scheme, but the reward is by no means near the levels we experience in hypothesis testing. Please refer to Section 2.4 in Chapter 2 for more details on the comparisons.

4.3.3 Exponential Distribution

In the third example, we consider samples that are distributed according to the exponential distribution. In particular, we assume that their density is

$$\mathsf{f}(x|\theta) = \theta e^{-\theta x},\ \theta > 0,\ x \geq 0,$$

while the prior is also exponential of the form

$$\pi(\theta) = a e^{-a\theta},\ a > 0,$$

where a is considered known. If we now compute the conditional pdf of θ given $\mathscr{F}_t$, then

$$\pi_t(\theta|\mathscr{F}_t) = \frac{S_t^{t+1}}{t!}\theta^t e^{-S_t\theta},\ \text{where } S_t = a + \textstyle\sum_{j=1}^t X_j,$$

which is Gamma$(\theta, t+1, S_t^{-1})$ distributed. From (4.10)

$$\hat{\vartheta}_t = \frac{h}{\tanh(\frac{hS_t}{t})}$$

and applying (4.11)

$$\mathscr{C}_t = 1 - \text{Gamma\_cdf}\big(\hat{\vartheta}_t + h, t+1, S_t^{-1}\big) + \text{Gamma\_cdf}\big(\hat{\vartheta}_t - h, t+1, S_t^{-1}\big).$$

It is interesting to note that the Bayesian estimator is not consistent since, using the law of large numbers (LLN), we have $\frac{S_t}{t} \to \frac{1}{\theta}$ a.s. and

$$\lim_{t\to\infty} \hat{\vartheta}_t = \frac{h}{\tanh(\frac{h}{\theta})} \neq \theta.$$

We can show that this limiting value has, in fact, error which is within the pre-specified

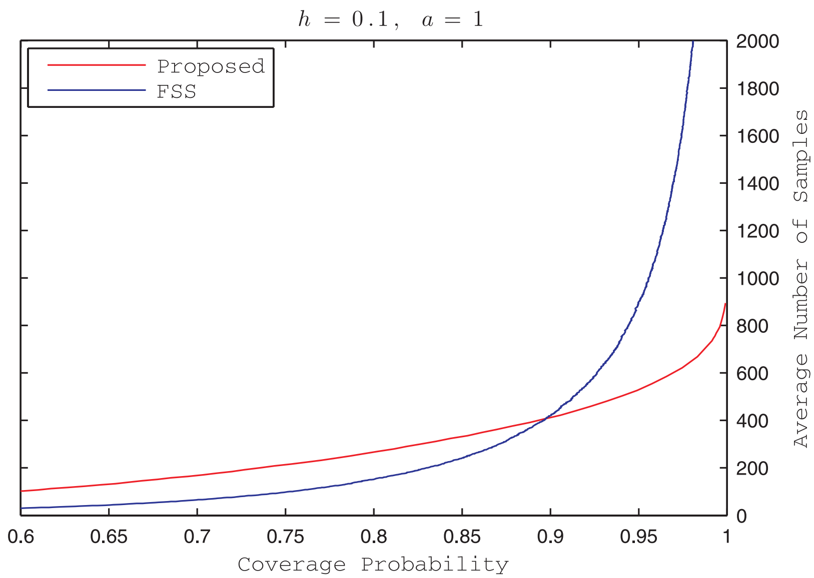


Figure 4.1: Average number of samples as a function of coverage probability for the exponential distribution when $h = 0.1$ and $a = 1$.

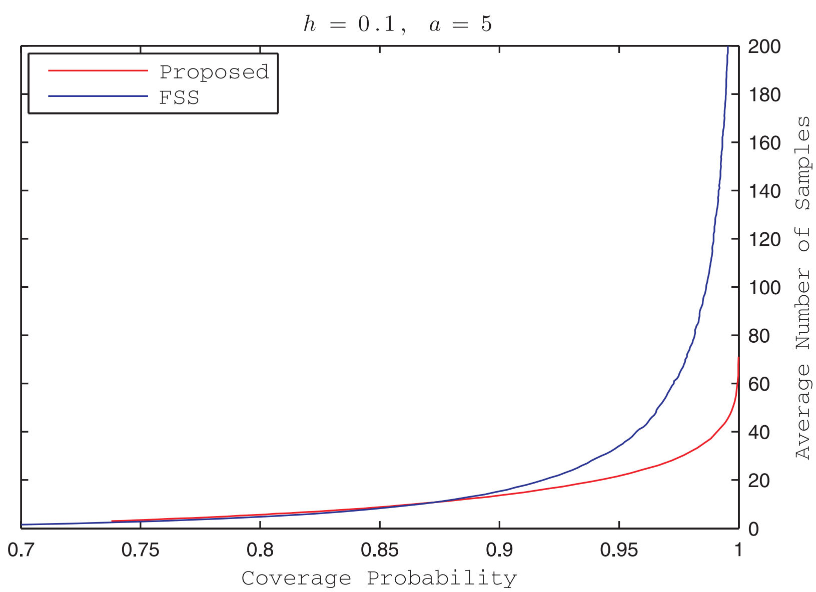


Figure 4.2: Average number of samples as a function of coverage probability for the exponential distribution when $h = 0.1$ and $a = 5$.

confidence interval since

$$\left|\frac{h}{\tanh(\frac{h}{\theta})} - \theta\right| \leq h.$$

From Fig. 4.3 and 4.4 we see that, for coverage values larger than 0.9 (which is the practically interesting range), the proposed method enjoys *substantial* gains as compared to the fixed-sample-size estimator. In particular, if the coverage probability is close to 0.99, the number of samples required by the proposed scheme is at least four times smaller than in the fixed-sample-size case. On the other hand, for coverage probabilities below 0.9 the fixed sample size prevails.

4.4 Conclusion

In the examples we presented, the performance of the proposed scheme was not always better than the fixed-sample-size estimator (although for high coverage probabilities it persistently outperformed it, and some times even considerably). This is because our method is not the solution of the optimization in (4.6). However, the proposed methodology is computationally efficient with good results. We are currently working on applying this methodology to the relative-width CI case, where problems of interval estimation of the normal mean and the exponential rate are of paramount importance in examples such as Markov Chain Monte Carlo (MCMC) simulations and queuing theory.

Acknowledgement

This work was supported by the US National Science Foundation under Grant CIF 1513373, through Rutgers University.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

In this thesis, we made three contributions to the sequential interval estimation research:

- Chapter 2 proposed an optimum fixed-width sequential methodology for the interval estimation of a binomial proportion when prior knowledge of the proportion is available. Regarding the optimum stopping time component, we demonstrated that it enjoys certain very interesting characteristics not commonly encountered in solutions of other classical optimal stopping problems. In particular, we proved that, for the Beta density prior, the optimum stopping time is always bounded from above and below; we need to first accumulate a sufficient amount of information before we start applying our stopping rule, and our stopping time will always terminate. We also conjectured that these properties are present with any prior. Furthermore, numerical performance evaluations showed that the proposed method exhibits an overall improved performance profile compared to its rivals. However, as discussed in Chapter 2, this optimum methodology is computationally expensive as we use backwards induction in dynamic programming.

- We proposed in Chapter 3 a simple but efficient methodology for scientists to use when no prior knowledge of the binomial proportion is available and when a computationally efficient method is desired. This methodology introduced the concept of tandem-width, where the half-width of the confidence interval of the proportion is not fixed beforehand; it is instead required to satisfy two different upper bounds depending on the values of the binomial proportion. To tackle this problem, we proposed a sequential method for obtaining fixed-width confidence intervals based on the minimax estimator of the binomial proportion. The tandem concept produced effective savings in sample size compared to the fixed-width counterpart. We suggest that scientists use this methodology as a first step when no prior knowledge of the true proportion is available rather than fixed-width or relative-width confidence intervals — otherwise, if the true proportion is close to zero, then it might take an unnecessarily large sample size to obtain a meaningful confidence interval for the proportion.

- In Chapter 4, we proposed an alternative formulation to the optimum one in Chapter 2 and extended our idea of sequential interval estimation to the case where we observe i.i.d. random variables with an arbitrary probability density function with unknown parameter of interest. For our analysis we adopted a conditional average cost approach that leads to a considerable simplification in the sequential estimation problem, otherwise known to be analytically intractable. Results showed that this methodology is superior to the optimum fixed-sample-size alternative.

We next present some goals that we aim to accomplish in the near future.

- We will extend the methodology in Chapters 2 and 4 to the relative-width confidence interval case, where the half-width h of the confidence interval is represented by $h = \eta\theta$ or $h = \eta\theta(1-\theta)$ for some unknown parameter θ to be estimated and $0 < \eta < 1$. The challenge with this extension of the material in Chapter 2 is to prove the nice properties of the optimum stopping time that are derived in Chapter 2. In particular, it is very challenging to prove that the optimum stopping time is bounded from above and below in the case of relative-width, even though our numerical results indicate that the stopping time is indeed bounded for the example of Beta(a, a) prior on θ for some $a > 0$. It is of interest to prove these properties for the relative case, as relative-width CI are important especially in the area of simulation. In Chapter 4, the relative case proves useful in interval estimation of a normal mean with unknown variance, as this is a hot topic in the simulation literature, since batch means are known to be (asymptotically) normally distributed.

- A second possible extension of our work is the sequential interval estimation of linear-regression coefficients. Here, the number of observations used for estimation is determined by the observed samples and hence is random, as opposed to fixed-sample-size estimation. Specifically, after receiving a new sample, if a target accuracy level is reached, we stop and provide the interval estimates using the samples collected so far; otherwise we continue to receive another sample.

REFERENCES

[1] N. Mukhopadhyay and B. de Silva, *Sequential Methods and Their Applications*. CRC Press, 2009.

[2] P. C. Mahalanobis, "A sample survey of the acreage under jute in Bengal," *Sankhya: The Indian Journal of Statistics (1933-1960)*, vol. 4, no. 4, pp. 511–530, 1940.

[3] A. Wald, *Sequential Analysis*. Wiley, 1947.

[4] A. Sullivan, D. Raben, J. Reekie, M. Rayment, A. Mocroft, *et al.*, "Feasibility and effectiveness of indicator condition-guided testing for HIV: Results from HIDES I (HIV indicator diseases across Europe study)," *PLoS ONE*, vol. 8, no. 1, e52845, 2013.

[5] J. Abramson, R. Takvorian, D. Fisher, Y. Feng, E. Jacobsen, *et al.*, "Oral clofarabine for relapsed/refractory non-Hodgkin lymphomas: Results of a phase 1 study.," *Leukemia & Lymphoma*, vol. 54, pp. 1915–1920, 2013.

[6] J. T. Morisette and S. Khorram, "Exact binomial confidence interval for proportions," *Photogrammetric Engineering & Remote Sensing*, vol. 64, no. 4, pp. 281–283, 1998.

[7] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, vol. 52, pp. 119–126, 1998.

[8] L. D. Brown, T. Cai, and A. Dasgupta, "Interval estimation for a binomial proportion (with discussion)," *Statistical Science*, vol. 16, pp. 101–133, 2001.

[9] R. G. Newcombe, "Two-sided confidence intervals for the single proportion: Comparison of seven methods," *Statistics in Medicine*, vol. 17, pp. 857–872, 1998.

[10] S. E. Vollset, "Confidence intervals for a binomial proportion," *Statistics in Medicine*, vol. 12, no. 9, pp. 809–824, 1993.

[11] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, pp. 209–212, 1927.

[12] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, pp. 404–413, 1934.

[13] T. E. Sterne, "Some remarks on confidence of fudicial limits," *Biometrika*, vol. 41, no. 1, pp. 275–278, 1954.

[14] E. L. Crow, "Confidence intervals for a proportion," *Biometrika*, vol. 43, pp. 423–435, 1956.

[15] C. R. Blyth and H. A. Still, "Binomial confidence intervals," *Journal of the American Statistical Association*, vol. 78, pp. 108–116, 1983.

[16] L. D. Brown, T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion and asymptotic expansions," *The Annals of Statistics*, vol. 30, pp. 160–201, 2002.

[17] J. Reiczigel, "Confidence intervals for the binomial parameter: Some new considerations," *Statistics in Medicine*, vol. 22, no. 4, pp. 611–621, 2003.

[18] A. M. Pires and C. Amado, "Interval estimators for a binomial proportion: Comparison of twenty methods," *REVSTAT - Statistical Journal*, vol. 6, no. 2, pp. 165–197, 2008.

[19] M. Thulin, "The cost of using exact confidence intervals for a binomial proportion," *Electronic Journal of Statistics*, vol. 8, no. 1, pp. 817–840, 2014.

[20] C. Jegourel, J. Sun, and J. S. Dong, "Sequential schemes for frequentist estimation of properties in Statistical Model Checking," in *Quantitative Evaluation of Systems: 14th International Conference, QEST 2017, Berlin, Germany, September 5-7, 2017, Proceedings*, N. Bertrand and L. Bortolussi, Eds. Springer International Publishing, 2017, pp. 333–350.

[21] Y. S. Chow and H. Robbins, "On the asymptotic theory of fixed-width sequential confidence intervals for the mean," *The Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 457–462, 1965.

[22] M. Tanaka, "On a confidence interval of given length for the parameter of the binomial and the Poisson distributions," *Annals of the Institute of Statistical Mathematics*, vol. 13, pp. 201–215, 1961.

[23] J. Frey, "Fixed-width sequential confidence intervals for a proportion," *The American Statistician*, vol. 64, pp. 242–249, 2010.

[24] P. Cabilio, "Sequential estimation in Bernoulli trials," *The Annals of Statistics*, vol. 5, no. 2, pp. 342–356, 1977.

[25] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer, 1988.

[26] A. N. Shiryaev, *Optimal Stopping Rules*. Springer, 1978.

[27] T. Choi and R. V. Ramamoorthi, "Remarks on consistency of posterior distributions," in *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, B. Clarke and S. Ghosal, Eds., ser. Collections. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2008, vol. Volume 3, pp. 170–186.

[28] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, 9th ed. Dover Publication, 1972.

[29] G. V. Moustakides, T. Yaacoub, and Y. Mei, "Sequential estimation based on conditional cost," *Proceedings of IEEE International Symposium on Information Theory*, pp. 436–440, 2017.

[30] R. A. Khan, "Fixed width confidence sequences for the normal mean and the binomial probability," *Sequential Analysis*, vol. 17, pp. 205–217, 1998.

[31] T. Yaacoub, G. V. Moustakides, and Y. Mei, *Optimal stopping for interval estimation in Bernoulli trials*. 2017. arXiv: `1711.06912`.

[32] E. L. Lehmann and G. Casella, *Theory of Point Estimation (2nd ed.)* New York: Springer-Verlag, 1998.

[33] D. Siegmund, *Sequential Analysis*. New York: Springer, 1985.

[34] M. A. Girshick, F. Mosteller, and L. J. Savage, "Unbiased estimates for certain binomial sampling problems with applications," *Annals of Mathematical Statistics*, vol. 17, pp. 13–23, 1946.

[35] J. R. Schultz, F. R. Nichol, G. L. Elfring, and S. D. Weed, "Multiple-stage procedures for drug screening," *Biometrics*, vol. 29, pp. 293–300, 1973.

[36] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[37] H. V. Poor, *An Introduction to Signal Detection and Estimation, 2nd edition*. New York: Springer, 1988.

[38] M. Ghosh, N. Mukhopadhyay, and P. K. Sen, *Sequential Estimation*. New York: Wiley, 1997.

[39] B. K. Ghosh, "On the attainment of the Cramer-Rao bound in the sequential case," *Sequential Analysis*, vol. 6, pp. 267–288, 1987.

[40] P. Grambsch, "Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator," *The Annals of Statistics*, vol. 11, pp. 68–77, 1983.

[41] G. Fellouris, "Asymptotically optimal parameter estimation under communication constraints," *The Annals of Statistics*, vol. 40, pp. 2239–2265, 2012.

[42] Y. Yilmaz, G. V. Moustakides, and X. Wang, "Sequential joint detection and estimation," *SIAM, Theory of Probability and Its Applications*, vol. 59, no. 3, pp. 452–465, 2015.

[43] Y. Yilmaz, S. Li, and X. Wang, "Sequential joint detection and estimation: Optimum tests and applications," *IEEE Transactions on Signal Processing*, vol. 64, no. 20, pp. 5311–5326, 2016.

[44] G. Casella, "Refining binomial confidence intervals," *Canadian Journal of Statistics*, vol. 14, pp. 113–129, 1986.

[45] P. Armitage, "Numerical studies in the sequential estimation of a binomial parameter," *Biometrika*, vol. 45, pp. 1–15, 1958.

[46] O. Marchal and J. Arbel, "On the sub-Gaussianity of the Beta and Dirichlet distributions," *Electronic Communications in Probability*, vol. 22, no. 54, pp. 1–14, 2017.